明細書

調査対象文書の文書特徴分析装置

技術分野

[0001] 本発明は、調査対象文書中の索引語の抽出に係わり、特に調査対象文書の性格乃至文書群に対する位置付けを的確に分析することを可能にする索引語の自動抽出装置、抽出プログラム、抽出方法、及び抽出された索引語を用いた性格表現図に関するものである。

また、本発明は文書の特徴分析装置に係わり、特に調査対象文書群に含まれる調査対象文書の、他の文書群に対する大まかな位置付けや、調査対象文書群全体としての特色を分析できるようにする文書の特徴分析装置、分析プログラム、分析方法、及び文書特徴表現図に関するものである。

背景技術

[0002] 特許文書をはじめ技術的文書やその他の文書は年々確実に量が増えている。近年、文書データが電子化されて流通するようになってから、膨大な文書群から調査対象の文書に類似した文書だけを自動検索するシステムが実用化されてきた。例えば、特開平11−73415号公報「類似文書検索装置及び類似文書検索方法」(特許文献1)においては、調査対象の文書に含まれる索引語を他の文書群に含まれる索引語と比較し、類似する索引語の種類や出現回数などから類似度を算出し、最も類似度の高い文書から順に出力している。

[0003] しかし、類似文書は検索されても、それだけでは調査対象の文書の性格或いは文書群での位置づけを知ることはできない。調査対象の文書の性格乃至文書群での位置づけを知るためには、検索結果の類似文書を読み込んだ上で、読み込んだ類似文書を前提として調査対象文書の評価をしなければならなかった。

[0004] 一方、文書の特徴そのものを自動抽出するものとして、例えば特開平11−345239号公報「文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体」(特許文献2)が存在する。この公報においては、「標準文書集合」から検索により「対象文書集合」を抽出し、この「対象文書集合」を構成する各「個別文書」の特

徴情報を抽出している。

具体的には、「対象文書集合」を「標準文書集合」に対して特徴付ける『対象文書集合全体特徴』を算出するとともに、「対象文書集合」中の各「個別文書」を他の個別文書に対して特徴付ける『個別文書特徴』を算出する。そして、これら『対象文書集合全体特徴』と『個別文書特徴』に基づいて、各「個別文書」の特徴情報を出力する。この技術は、大量の情報の中からユーザが有益な情報を見つけ出して取捨選択することを容易にする点で有益である。

特許文献1:特開平11−73415号公報「類似文書検索装置及び類似文書検索方法」

特許文献2:特開平11−345239号公報「文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体」

発明の開示

発明が解決しようとする課題

[0005] しかし、上記特開平11−345239号公報(特許文献2)に記載の技術には、次の3つの問題がある。

[0006] 第1に、この公報に記載の技術では例えば「桜の花見」など特定のテーマを決めてからこれに合致する「対象文書集合」を抽出する。そしてこの「対象文書集合」が抽出されることで初めて、特徴情報の抽出対象となる各「個別文書」が決定される。すなわち、「対象文書集合」やそれを抽出する特定のテーマが予め決まっていないと「個別文書」を決定することさえできない。従ってこの公報に記載の技術では、特定の調査対象文書が与えられたときにその性格を分析することはできない。

第2に、この公報に記載の技術では『対象文書集合全体特徴』と『個別文書特徴』との積を算出することで、「対象文書集合」を特徴付け且つ各「個別文書」を特徴付ける情報として出力する。従ってこの公報に記載の技術では、特徴情報を単に1次元的な量で捉えるにとどまり、調査対象文書の性格を多面的に分析することはできない。

第3に、調査対象文書群に含まれる調査対象文書の、他の文書群に対する大まかな位置付けや、調査対象文書群全体としての傾向を、専門性や独創性といった観点

から分析することのできる文書の特徴分析装置は開示されておらず、他の文献にも記載されていない。

[0007] 本発明の第1の課題は、調査対象文書が与えられたときにその性格の的確な把握を可能にする索引語抽出装置を提供することである。

また本発明の第2の課題は、調査対象文書の性格の多面的な分析を可能にする索引語抽出装置及び性格表現図を提供することである。

また本発明の第3の課題は、調査対象文書群に含まれる調査対象文書の、他の文書群に対する大まかな位置付けや、調査対象文書群全体としての傾向を分析することを可能にする文書特徴分析装置及び文書特徴表現図を提供することである。

課題を解決するための手段

[0008] 上記第1の課題を解決するため、本発明の索引語抽出装置は、調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群の選出元となる選出源文書群、を入力する入力手段と、前記調査対象文書内の索引語を抽出する索引語抽出手段と、前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出手段と、前記調査対象文書のデータに基づき、前記選出源文書群の中から前記類似文書群を選出する類似文書群選出手段と、前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出手段と、各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記類似文書群における出現頻度の関数値との組合せに基づき、各索引語とその位置づけデータとを出力する出力手段と、を備えている。

本発明は、出現頻度の関数値を各索引語における組合せで観察することにより、調査対象文書の性格を分析できるようにするものである。

本発明によれば、調査対象文書内の索引語を抽出する処理、選出源文書群から類似文書群を選出する処理、比較対象文書群又は類似文書群における出現頻度の関数値を算出する処理等は、すべてコンピュータで行われるので、各処理のために人間が文書内容を読む必要はまったくない。

特に、調査対象文書のデータに基づいて類似文書群を新たに選出し、この類似文

書群における出現頻度の関数値と、比較対象文書群における出現頻度の関数値との組合せに基づき、各索引語とその位置づけデータとを出力するので、調査対象文書の性格を精度よく分析することができる。

上記比較対象文書群及び選出源文書群は、検索処理可能なデータである必要はあるが、内容については格別の制約はなく、例えばこれらが同一の文書群であっても良いし、異なる文書群であってもよい。また、これらの文書群の何れか又は双方が、ある文書群から無作為抽出されたものでも良いし、一定条件のもとで全件抽出されたものでもよい。典型例としては、ある国及び期間における全特許文書(公開特許公報など)を、比較対象文書群及び選出源文書群とする。

上記調査対象文書は、1文書でも複数の文書でもよい。複数の文書をまとめて調査対象文書とする場合は、個々の調査対象文書の性格というよりは、文書群としての性格を示すことになる。また調査対象文書は、比較対象文書群又は選出源文書群に含まれるものでも、含まれないものでもよい。

上記索引語抽出手段による索引語の抽出は、文書の全部又は一部から単語を切り出すことにより行う。単語の切り出し方に特段の制約はなく、例えば日本語文書であれば従来から知られている方法や市販の形態素解析ソフトを活用して、助詞や接続詞を除き、意味ある品詞を抽出する方法でも良いし、索引語の辞書(シソーラス)のデータベースを事前に保持し、該データベースから得られる索引語を利用する方法でもよい。

索引語の文書群における出現頻度としては、例えば、当該文書群を検索対象とし、ある索引語で検索したときのヒット文書数(文書頻度DF)を用いるが、これに限られるものではなく、例えば当該索引語がヒットした延べ回数でもよい。

出力手段による索引語の出力は、索引語抽出手段により抽出された索引語すべてを出力しても良いし、文書の性格を強く示す一部の索引語のみを出力しても良い。また、出力手段により索引語とともに出力される位置づけデータは、比較対象文書群及び類似文書群における出現頻度の関数値をそのままの形で出力しても良いし、これに基づいて座標上に索引語を配置した図として出力しても良いし、上記出現頻度の関数値に基づいてグループ分けされた索引語のリストとして出力しても良い。

[0009] 上記索引語抽出装置においては、前記選出源文書群として前記比較対象文書群を用いることとするのが好ましい。これにより選出源文書群の入力を、比較対象文書群の入力と別々にする必要がなくなり、構成を簡略化することができる。また、類似文書群が比較対象文書群の部分集合になるので、データの解析がより容易になる。

[0010] 上記索引語抽出装置において、前記類似文書群選出手段は、前記調査対象文書及び前記選出源文書群の各文書について、当該文書に含まれる各索引語の当該文書における出現頻度の関数値又は各索引語の前記選出源文書群における出現頻度の関数値を成分とするベクトルを算出し、前記調査対象文書について算出された前記ベクトルに対する類似度合いの高いベクトルをもつ文書を前記選出源文書群から選出して、類似文書群とすることが望ましい。

類似文書群の選出を各文書のベクトルに基づいて行うので、高い信頼性を確保することができる。また、例えばIPC(国際特許分類)等の一致により類似文書群を選出する場合と異なり、類似度合いの高い順に何件という形での件数指定も自在にできる。

上記ベクトルの類似度合いの判定は、ベクトル間の余弦乃至Tanimoto相関(類似度)などベクトル成分間の積の関数を用いても良いし、ベクトル間の距離(非類似度)などベクトル成分間の差の関数を用いてもよい。

[0011] 上記索引語抽出装置において、前記出力手段は、各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の低い第1グループの索引語と、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの索引語と、前記第1グループの索引語よりも前記類似文書群における出現頻度が高い第3グループの索引語と、をそれぞれ出力することが望ましい。

比較対象文書群における出現頻度の関数値と、類似文書群における出現頻度の関数値とを用いて、第1ー第3グループの索引語をそれぞれ出力することにより、調査対象文書の性格を多面的に分析することができる。

例えば上記第1グループの索引語は、調査対象文書に含まれる専門的な内容、又はこれに直結する概念を表現する語(専門語)であると評価できる。

また例えばここでいう第2グループには、比較対象文書群における出現頻度の高さの割には、類似分野ではあまり着目されていなかった概念を示す語(独創的着目語)が含まれる。

また例えばここでいう第3グループには、類似文書群を特徴付ける語(類似文書群規定語)が含まれる。例えば技術文書を調査対象とした場合であれば、この第3グループの索引語を見れば、類似文書群及び調査対象文書の技術分野を知ることができる。

[0012] 上記索引語抽出装置において、前記出力手段は、各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の高い第4グループの索引語よりも、前記比較対象文書群における出現頻度が低い第3グループの索引語と、前記第4グループの索引語よりも前記類似文書群における出現頻度が低い第2グループの索引語と、前記第3グループの索引語よりも前記類似文書群における出現頻度が低く且つ前記第2グループの索引語よりも前記比較対象文書群における出現頻度が低い第1グループの索引語と、をそれぞれ出力することが望ましい。

比較対象文書群における出現頻度の関数値と、類似文書群における出現頻度の関数値とを用いて、第1ー第3グループの索引語をそれぞれ出力することにより、調査対象文書の性格を多面的に分析することができる。

例えばここでいう第3グループの索引語は、類似文書群を特徴付ける語(類似文書群規定語)であると評価できる。例えば技術文書を調査対象とした場合であれば、この第3グループの索引語を見れば、類似文書群及び調査対象文書の技術分野を知ることができる。

また例えばここでいう第2グループの索引語は、比較対象文書群における出現頻度の高さの割には、類似分野ではあまり着目されていなかった概念を示す語(独創的着目語)であると評価できる。

また例えば上記第1グループの索引語は、調査対象文書に含まれる専門的な内容、又はこれに直結する概念を表現する語(専門語)であると評価できる。

ここでいう第3グループや第2グループには、前記比較対象文書群においても前記

類似文書群においても出現頻度の高い第4グループの索引語(一般語)は含まれないので、精度の高い分析が可能である。

[0013] 上記第2の課題を解決するため、本発明の索引語抽出装置は、調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群、を入力する入力手段と、前記調査対象文書内の索引語を抽出する索引語抽出手段と、前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出手段と、前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出手段と、各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の低い第1グループの索引語と、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの索引語と、前記第1グループの索引語よりも前記類似文書群における出現頻度が高い第3グループの索引語と、をそれぞれ出力する出力手段と、を備えている。

調査対象文書内の索引語の、比較対象文書群における出現頻度の関数値と、類似文書群における出現頻度の関数値と、に基づいて、第1ー第3グループの索引語をそれぞれ出力することにより、調査対象文書の性格を多面的に分析することができる。

例えば上記第1グループの索引語は、調査対象文書に含まれる専門的な内容、又はこれに直結する概念を表現する語(専門語)であると評価できる。

また例えばここでいう第2グループには、比較対象文書群における出現頻度の高さの割には、類似分野ではあまり着目されていなかった概念を示す語(独創的着目語)が含まれる。

また例えばここでいう第3グループには、類似文書群を特徴付ける語(類似文書群規定語)が含まれる。例えば技術文書を調査対象とした場合であれば、この第3グループの索引語を見れば、類似文書群及び調査対象文書の技術分野を知ることができる。

本発明によれば、調査対象文書内の索引語を抽出する処理、比較対象文書群又は類似文書群における出現頻度の関数値を算出する処理等は、すべてコンピュータ

で行われるので、各処理のために人間が文書内容を読む必要はまったくない。

比較対象文書群は、検索処理可能なデータである必要はあるが、それ以外に格別の制約はなく、例えばある文書群から無作為抽出されたものでも良いし、一定条件のもとで全件抽出されたものでもよい。例えば、ある国及び期間における全特許文書(公開特許公報など)を、比較対象文書群とする。

類似文書群も、検索処理可能なデータである必要がある。類似文書群は調査対象文書のデータに基づいて比較対象文書群などの文書群から選出して入力しても良いし、調査対象文書のデータに基づかないで選出したものを入力しても良い。例えば、公知の方法により選出した類似文書群の中から調査対象文書を選んでこれらを入力することにより、結果として当該類似文書群が調査対象文書に類似する類似文書群となる場合でもよい。

調査対象文書は、1文書でも複数の文書でもよい。複数の文書をまとめて調査対象文書とする場合は、個々の調査対象文書の性格というよりは、文書群としての性格を示すことになる。また調査対象文書は、比較対象文書群又は類似文書群に含まれるものでも、含まれないものでもよい。

索引語抽出手段による索引語の抽出は、文書の全部又は一部から単語を切り出すことにより行う。単語の切り出し方に特段の制約はなく、例えば日本語文書であれば従来から知られている方法や市販の形態素解析ソフトを活用して、助詞や接続詞を除き、意味ある品詞を抽出する方法でも良いし、索引語の辞書(シソーラス)のデータベースを事前に保持し、該データベースから得られる索引語を利用する方法でもよい。

索引語の文書群における出現頻度としては、例えば、当該文書群を検索対象とし、ある索引語で検索したときのヒット文書数(文書頻度DF)を用いるが、これに限られるものではなく、当該索引語がヒットした延べ回数でもよい。

出力手段による索引語の出力は、索引語抽出手段により抽出された索引語すべてを位置づけデータとともに出力しても良いし、文書の性格を良く示す一部の索引語のみを出力しても良い。

[0014] また、本発明の索引語抽出装置は、調査対象文書、前記調査対象文書と比較され

る比較対象文書群、前記調査対象文書に類似する類似文書群、を入力する入力手段と、前記調査対象文書内の索引語を抽出する索引語抽出手段と、前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出手段と、前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出手段と、各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の高い第4グループの索引語よりも、前記比較対象文書群における出現頻度が低い第3グループの索引語と、前記第4グループの索引語よりも前記類似文書群における出現頻度が低い第2グループの索引語と、前記第3グループの索引語よりも前記類似文書群における出現頻度が低く且つ前記第2グループの索引語よりも前記比較対象文書群における出現頻度が低い第1グループの索引語と、をそれぞれ出力する出力手段と、を備えている。

調査対象文書内の索引語の、比較対象文書群における出現頻度の関数値と、類似文書群における出現頻度の関数値と、に基づいて、第1ー第3グループの索引語をそれぞれ出力することにより、調査対象文書の性格を多面的に分析することができる。

例えばここでいう第3グループの索引語は、類似文書群を特徴付ける語(類似文書群規定語)であると評価できる。例えば技術文書を調査対象とした場合であれば、この第3グループの索引語を見れば、類似文書群及び調査対象文書の技術分野を知ることができる。

また例えばここでいう第2グループの索引語は、比較対象文書群における出現頻度の高さの割には、類似分野ではあまり着目されていなかった概念を示す語(独創的着目語)であると評価できる。

また例えば上記第1グループの索引語は、調査対象文書に含まれる専門的な内容、又はこれに直結する概念を表現する語(専門語)であると評価できる。

ここでいう第3グループや第2グループには、前記比較対象文書群においても前記類似文書群においても出現頻度の高い第4グループの索引語(一般語)は含まれないので、精度の高い分析が可能である。

[0015] 上記各索引語抽出装置において、前記比較対象文書群又は前記類似文書群における出現頻度の関数値は、当該出現頻度の逆数に、前記比較対象文書群又は前記類似文書群の総文書数を乗じたものの対数であることが望ましい。

これにより、出現頻度の関数値が特定の値付近に集中することを避け、索引語の位置づけの把握を容易にすることができる。特に各索引語を座標上に配置した場合には、各索引語が原点付近に集中することを避け、位置づけの視覚的な把握を容易にすることができる。

[0016] 上記各索引語抽出装置において、前記出力手段は、前記比較対象文書群における出現頻度の関数値を座標の第1軸にとり、前記類似文書群における出現頻度の関数値を前記座標の第2軸にとって、前記索引語を配置し出力することが望ましい。

座標上に配置された索引語の位置により、各索引語の位置づけを視覚的に把握することができる。すなわち、座標上の2次元的な配置により、上記第1ー第3グループの索引語の区別を一見して明瞭に把握することができる。

座標系としては例えば平面の直交座標を用い、第1軸としてX軸(横軸)、第2軸としてY軸(縦軸)を用いるが、これに限らず例えば3次元座標を用いて上記以外の指標をZ軸にとってもよい。

[0017] 上記各索引語抽出装置において、前記出力手段は、前記第1グループの索引語と、前記第2グループの索引語と、前記第3グループの索引語とを、それぞれリストして出力することが望ましい。

これにより、各領域に属する索引語をリストの状態で見ることができる。このリストは例えば各文書群における出現頻度に応じた順序で索引語をソートしたものとすることにより、調査対象文書の性格分析をより的確に行うことができる。

[0018] 上記各索引語抽出装置において、前記出力手段は、前記第1グループの索引語と、前記第2グループの索引語と、前記第3グループの索引語とを用いて、当該調査対象文書の解説文を自動生成して出力することが望ましい。

これにより、調査対象文書の性格を述べる解説文として出力することができる。この解説文は、例えば、「**、**(第3グループの索引語)に関する技術分野において、**、**(第1グループの索引語)に関わる専門的な概念・技術を利用し、**、**(第2グ

ループの索引語）の観点に着目した文書」のように生成する。

また例えば第1グループに該当する索引語が存在しないときは、解説文は第1グループの索引語に関する記述を除き、「**、**（第3グループの索引語）に関する技術分野において、**、**（第2グループの索引語）の観点に着目した文書」のように生成する。

[0019] 上記各索引語抽出装置において、前記類似文書群の各文書は、前記比較対象文書群に含まれており、前記出力手段は、前記比較対象文書群における出現頻度の関数値を、さらに変換して座標の第1軸にとり、前記類似文書群における出現頻度の関数値を前記座標の第2軸にとって、前記索引語を配置し出力するものであって、前記変換は、前記類似文書群が前記比較対象文書群の部分集合であることによる、前記索引語の前記座標上における存在可能領域の境界線が、前記第1軸と垂直に近づくように変換することが望ましい。

類似文書群を選出するための選出源文書群を比較対象文書群とした場合には、類似文書群は比較対象文書群の部分集合となる。従って、例えばある索引語を比較対象文書群Pで検索したときのヒット文書数DF（P）は、同じ索引語を類似文書群Sで検索したときのヒット文書数DF（S）より小さい数にはなり得ない。従って例えば上記DF（P）を直交座標のX軸に、上記DF（S）をY軸にとろうとすると、X≧Yの領域にのみ各索引語が配置されることになるので、存在可能領域の境界線が45度に傾いた状態となる。また例えば上記DF（P）の逆数に比較対象文書総数Nを乗じたものの対数IDF（P）を直交座標のX軸に、上記DF（S）の逆数に類似文書総数N′を乗じたものの対数IDF（S）をY軸にとろうとすると、Y≧X−ln（N／N′）の領域（ここでは対数として自然対数を用いた）にのみ各索引語が配置されることになるので、存在可能領域の境界線が45度に傾いた状態となる。

本発明によれば、各索引語を座標上に配置した場合の存在可能領域が矩形に近づくので、各索引語がどの領域に属するかの視覚的把握を一層容易にすることができる。

[0020] 上記索引語抽出装置において、前記変換は、前記類似文書群における出現頻度との関数によって与えられる変換であることが望ましい。

例えば、変換前の点の座標を(X, Y)とおいた場合、変換後の点の座標(X′ , Y′)＝(X−Y＋const, Y)とする。また例えば、変換後の点の座標(X′ , Y′)＝(X＊($\alpha + \beta_2/2$)／(Y＋α), Y)とする。

これにより、索引語座標の存在可能領域を矩形に近づける際に、索引語座標の横軸に沿った移動量が縦軸の値によって異なるようにし、原点付近などへの索引語座標の集中を避けることができる。

[0021] 上記各索引語抽出装置において、前記調査対象文書内の各索引語の、当該調査対象文書における出現頻度を算出する索引語頻度算出手段を更に備え、前記出力手段は、前記調査対象文書内の各索引語の当該調査対象文書における出現頻度を反映して出力することが望ましい。

これにより、調査対象文書における各索引語の重みを加味して調査対象文書の性格を分析することができる。

反映のさせ方としては、例えば、比較対象文書群又は類似文書群における出現頻度の関数値に基づいて各索引語を座標に配置する場合には、調査対象文書内の各索引語の当該調査対象文書における出現頻度(TF)の大小によって異なる色を用いて各索引語を表示する方法、各索引語の出現頻度(TF)をZ成分とし、3次元グラフィックにより3次元座標を表示する方法、等が考えられる。また例えばいわゆるTFIDFを用いて、各索引語の位置づけデータを出力する方法が考えられる。

なお、索引語頻度算出手段により算出された調査対象文書内の各索引語の出現頻度は、類似文書群を選出する場合の文書の類似度合いの判定にも用いることができる。

[0022] 上記各索引語抽出装置において、前記出力手段は、各索引語につき、前記比較対象文書群における出現頻度の関数値を座標の第1軸にとり、前記類似文書群における出現頻度の関数値を前記座標の第2軸にとった場合に、前記座標上の複数の基準点のうち当該索引語に最も近い基準点に更に近づくように配置して座標上に出力することが望ましい。

これにより、索引語の位置が基準点に近づくので、座標上の表示をより見易くすることができる。このような処理のためには、自己組織化マップ(SOM)を応用した技術を

用いることが望ましい。

[0023] 上記各索引語抽出装置において、座標上に複数の基準点の座標を設定する基準点設定手段と、各索引語につき、前記比較対象文書群における出現頻度の関数値を座標の第1軸にとり、前記類似文書群における出現頻度の関数値を前記座標の第2軸にとった場合に、前記複数の基準点のうち当該索引語に最も近い基準点の座標データを、当該索引語に更に近づくように、所定回数にわたり更新する手段と、前記更新された基準点に基づいて、当該索引語を配置する座標を算出する座標算出手段と、を更に備え、前記出力手段は、前記座標算出手段により算出された座標に基づいて、各索引語を前記座標に配置して出力することが望ましい。

これにより、索引語の位置が基準点に近づくので、座標上の表示をより見易くすることができる。

[0024] 本発明の性格表現図は、調査対象文書内の索引語について、前記調査対象文書と比較される比較対象文書群における出現頻度の関数値を座標の第1軸にとり、前記調査対象文書に類似する類似文書群における出現頻度の関数値を前記座標の第2軸にとって配置したものである。

座標上に配置された索引語の位置により、各索引語の位置づけを視覚的に把握できる結果、調査対象文書の性格を的確に分析することができる。すなわち、座標上の2次元的な配置により、上記第1ー第3グループの索引語の区別を一見して明瞭に把握することができる。

座標系としては例えば平面の直交座標を用い、第1軸としてX軸(横軸)、第2軸としてY軸(縦軸)を用いるが、これに限らず例えば3次元座標を用いて上記以外の指標をZ軸にとってもよい。

[0025] 本発明の他の性格表現図は、調査対象文書内の索引語を配置した、調査対象文書の性格表現図であって、第1エリアに、前記調査対象文書と比較される比較対象文書群においても、前記調査対象文書群に類似する類似文書群においても、出現頻度の低い第1グループの索引語を配置し、第2エリアに、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの索引語を配置し、第3エリアに、前記第1グループの索引語よりも前記類似文書群における出現

頻度が高い第3グループの索引語を配置したものである。

出現頻度の関数値に基づいて、第1エリアー第3エリアに各索引語を配置することにより、調査対象文書の性格を多面的に分析することができる。

例えば上記第1グループの索引語は、調査対象文書に含まれる専門的な内容、又はこれに直結する概念を表現する語(専門語)であると評価できる。

また例えばここでいう第2エリアには、比較対象文書群における出現頻度の高さの割には、類似分野ではあまり着目されていなかった概念を示す語(独創的着目語)が含まれる。

また例えばここでいう第3グループには、類似文書群を特徴付ける語(類似文書群規定語)が含まれる。例えば技術文書を調査対象とした場合であれば、この第3グループの索引語を見れば、類似文書群及び調査対象文書の技術分野を知ることができる。

この性格表現図は、2次元座標上に索引語を配置したものでも良いし、索引語を列挙する表の各欄を各エリアに割り当てて索引語を表示したものでも良い。

[0026] 本発明の他の性格表現図は、調査対象文書内の索引語を配置した、調査対象文書の性格表現図であって、第3エリアに、前記調査対象文書と比較される比較対象文書群においても前記調査対象文書群に類似する類似文書群においても出現頻度の高い第4グループの索引語よりも、前記比較対象文書群における出現頻度が低い第3グループの索引語を配置し、第2エリアに、前記第4グループの索引語よりも前記類似文書群における出現頻度が低い第2グループの索引語を配置し、第1エリアに、前記第3グループの索引語よりも前記類似文書群における出現頻度が低く且つ前記第2グループの索引語よりも前記比較対象文書群における出現頻度が低い第1グループの索引語を配置したものである。

出現頻度の関数値に基づいて、第1エリアー第3エリアに各索引語を配置することにより、調査対象文書の性格を多面的に分析することができる。

例えばここでいう第3グループの索引語は、類似文書群を特徴付ける語(類似文書群規定語)であると評価できる。例えば技術文書を調査対象とした場合であれば、この第3グループの索引語を見れば、類似文書群及び調査対象文書の技術分野を知

ることができる。

また例えばここでいう第2グループの索引語は、比較対象文書群における出現頻度の高さの割には、類似分野ではあまり着目されていなかった概念を示す語(独創的着目語)であると評価できる。

また例えば上記第1グループの索引語は、調査対象文書に含まれる専門的な内容、又はこれに直結する概念を表現する語(専門語)であると評価できる。

ここでいう第3グループや第2グループには、前記比較対象文書群においても前記類似文書群においても出現頻度の高い第4グループの索引語(一般語)は含まれないので、精度の高い分析が可能である。

[0027] 上記第3の課題を解決するため、本発明の文書特徴分析装置は、複数の調査対象文書を含む調査対象文書群、各調査対象文書と比較される比較対象文書群、前記調査対象文書群と共通の属性を有する同類文書群、を入力する入力手段と、前記各調査対象文書内の索引語を抽出する索引語抽出手段と、前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値、を算出する第3出現頻度算出手段と、前記抽出された索引語の、前記同類文書群における出現頻度の関数値、を算出する第4出現頻度算出手段と、各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記同類文書群における出現頻度の関数値との組合せに基づき、前記調査対象企業の各文書における中心点を算出する中心点算出手段と、前記各調査対象文書における前記中心点のデータを出力する出力手段と、を備えている。

これにより、調査対象文書群に含まれる調査対象文書について、比較対象文書群及び同類文書群に対する大まかな位置付けを知ることができる。例えば、比較対象文書群及び同類文書群に対して、標準的文書なのか、独創的性質を持つ文書か、或いは専門的性質を持つ文書かを知ることができる。また例えば、調査対象文書群から、標準的文書、独創的性質を持つ文書、又は専門的性質を持つ文書を検出することができる。更に、調査対象文書群全体としての傾向を評価するこができる。例えば、標準的文書の多い文書群、独創的性質を持つ文書の多い文書群、或いは専門的性質を持つ文書の多い文書群というように評価することができる。

上記調査対象文書群は、例えばある調査対象企業の文書群、或いは調査対象技術分野の文書群などが挙げられる。前者の場合は例えば全特許文書群から調査対象企業を出願人とする文書をすべて検索し、或いは更にIPC等で絞り込んで調査対象文書群とする。後者の場合は例えば全特許文書群から特定のIPCが付与された文書をすべて検索し、或いは更に出願期間等で絞り込んで調査対象文書群とする。上記調査対象文書群は、比較対象文書群及び同類文書群に含まれるものであることが望ましいが、含まれないものでもよい。

上記比較対象文書群は、検索処理可能なデータである必要はあるが、内容については格別の制約はなく、例えばある文書群から無作為抽出されたものでも良いし、一定条件のもとで全件抽出されたものでもよい。典型例としては、ある国及び期間における全特許文書(公開特許公報など)を、比較対象文書群とする。

上記同類文書群も、検索処理可能なデータである必要はあるが、その選出方法に格別の制約はない。例えば調査対象文書群を調査対象企業の文書群とする場合には、同類文書群は、ユーザが当該調査対象企業と同業界の企業名を複数指定して検索された文書群でもよいし、調査対象企業の企業名と業界分類から同業界の企業を検索するようにしても良い。また、調査対象企業の文書と同分野に属する文書群をIPC(国際特許分類)などにより検索するようにしても良い。また、これら同業界の文書群又は同分野の文書群から更に一定条件で絞り込んでも良い。

また、例えば調査対象文書群を調査対象技術分野の文書群とする場合には、(例えばIPCのサブグループまで指定して検索した)特定の技術分野に属する調査対象文書群より広い範囲の技術分野に含まれる文書群を、(例えばIPCのメイングループまでの指定で検索し)同類文書群とする。また、例えばIPCで検索し更に特定の出願期間で絞り込んだ調査対象文書群より、長い出願期間で絞り込んで同類文書群とする。

同類文書群は、比較対象文書群の中から選出することが望ましいがこれに限られるものではない。調査対象企業の文書をIPCで絞り込んだ文書群を上記調査対象文書群とする場合は、同類文書群も同じIPCで検索し或いは絞り込んだものを使うことが好ましい。

上記索引語抽出手段による索引語の抽出は、文書の全部又は一部から単語を切り出すことにより行う。単語の切り出し方に特段の制約はなく、例えば日本語文書であれば従来から知られている方法や市販の形態素解析ソフトを活用して、助詞や接続詞を除き、意味ある品詞を抽出する方法でも良いし、索引語の辞書(シソーラス)のデータベースを事前に保持し、該データベースから得られる索引語を利用する方法でもよい。

索引語の文書群における出現頻度としては、例えば、当該文書群を検索対象とし、ある索引語で検索したときのヒット文書数(文書頻度DF)を用いるが、これに限られるものではなく、例えば当該索引語がヒットした延べ回数でもよい。

また、出現頻度の関数値としては、当該出現頻度の逆数に、前記比較対象文書群又は前記同類文書群の総文書数を乗じたものの対数(IDF)であることが望ましい。

上記各調査対象文書における中心点は、例えば座標(<IDF(P)>$_w$, <IDF(S)>$_w$)で与えられる点(但し"< >$_w$"は各文書における平均値)とするが、これに限られるものではない。

上記出力手段は、上記中心点を座標上に配置したマップとして出力することが望ましい。座標系としては例えば平面の直交座標を用い、第1軸としてX軸(横軸)、第2軸としてY軸(縦軸)を用いるが、これに限らず例えば3次元座標を用いて上記以外の指標をZ軸にとってもよい。

[0028] 上記文書特徴分析装置において、各調査対象文書における前記中心点の算出は、各索引語についての、前記比較対象文書群における出現頻度の関数値と前記同類文書群における出現頻度の関数値とに基づく各索引語の座標値に、当該文書内の索引語頻度値合計に対する各索引語の索引語頻度値の比で重み付けをした平均値である索引語座標の加重平均値を算出することによって行うことが望ましい。

これにより、中心点の算出に索引語頻度による重み付けを反映させることができる。

[0029] 上記文書特徴分析装置において、前記調査対象文書群のうち、当該文書群に対して類似性の高い文書と、当該文書群に対して類似性の低い文書とを抽出して前記中心点のデータを出力することが望ましい。

調査対象文書群の文書数が膨大にある場合でも、代表的な文書に絞って出力する

ことにより、調査対象文書群としての傾向を把握し易くすることができる。

調査対象文書群に対する各文書の類似性の判定は、例えば、各文書dにつき、各索引語w_iで調査対象文書群(E0)内を検索したときのヒット文書数DF(w_i, E0)の平均値(1／d_N){DF(w_1, E0)＋DF(w_2, E0)＋・・・＋DF(w_{dN}, E0)}の高いものを「類似」、低いものを「非類似」とする(d_Nは当該文書d内の索引語数)。抽出の方法としては、例えば上記平均値の昇順及び降順の一定数を抽出する方法、また例えば上記平均値を企業内文書数で除したものをZとしたときに、「全Zの平均値＋全Zの標準偏差」以上のZをとる文書と、「全Zの平均値−全Zの標準偏差」以下のZをとる文書とを抽出する方法などが考えられる。

[0030] 本発明の、調査対象文書の文書特徴表現図は、調査対象文書群に含まれる複数の調査対象文書について、各調査対象文書と比較される比較対象文書群に対する位置づけを座標の第1軸にとり、前記調査対象文書群と共通の属性を有する同類文書群に対する位置づけを前記座標の第2軸にとって配置したものであって、前記座標における前記各調査対象文書の座標値は、各調査対象文書内の各索引語の前記比較対象文書群における出現頻度の関数値と、各索引語の前記同類文書群における出現頻度の関数値と、を成分とする索引語座標値の、各調査対象文書における中心点としたものである。

これにより、調査対象文書群全体の傾向を分析することができる。

上記調査対象文書群の各文書における中心点は、例えば座標(＜IDF(P)＞$_w$, ＜IDF(S)＞$_w$)で与えられる点(但し"＜　＞$_w$"は各文書における平均値)とするが、これに限られるものではない。また例えば、当該調査対象文書内の索引語頻度値合計に対する各索引語の索引語頻度値の比で重み付けをした平均値であってもよい。

[0031] また本発明は、上記各装置によって実行される方法と同じ工程を備えた抽出方法及び分析方法、並びに上記各装置によって実行される処理と同じ処理をコンピュータに実行させることのできる抽出プログラム及び分析プログラムである。このプログラムは、FD、CDROM、DVDなどの記録媒体に記録されたものでもよく、ネットワークで送受信されるものでもよい。

発明の効果

[0032] 本発明によれば、第1に、調査対象文書が与えられたときにその性格を的確に表現できるようにする索引語抽出装置を提供することができる。

また第2に、調査対象文書の性格を多面的に分析できるようにする索引語抽出装置及び性格表現図を提供することができる。

また第3に、調査対象文書群に含まれる調査対象文書の、他の文書群に対する大まかな位置付けや、調査対象文書群全体としての傾向を分析できるようにする文書特徴分析装置及び文書特徴表現図を提供することができる。

図面の簡単な説明

[0033] [図1]本発明の一実施形態に係る特徴索引語抽出装置のハードウェア構成を示す図。

[図2]上記特徴索引語抽出装置における構成と機能を詳細に説明する図。

[図3]入力装置2における条件設定の動作を示すフローチャート。

[図4]処理装置1の動作を示すフローチャート。

[図5]出力装置4におけるマップ、リスト、及びコメントの出力の動作を示すフローチャート。

[図6]調査対象文書の入力条件設定画面の表示例を示す図。

[図7]比較対象文書群の入力条件設定画面の表示例を示す図。

[図8]索引語抽出条件および類似文書群選出条件の設定画面の表示例を示す図。

[図9]出力条件設定画面の表示例を示す図。

[図10]マップの性質を説明するための概念図。

[図11]実施例1の特徴索引語抽出装置による「外部補助記憶装置」に関する公開特許公報のマップ表示の具体例を示す図。

[図12]図11と同じ調査対象文書に関する、リスト出力の具体例を示す図。

[図13]実施例1の特徴索引語抽出装置による「緊急通報」に関する公開特許公報のマップ表示の具体例を示す図。

[図14]図13と同じ調査対象文書に関する、リスト出力の具体例を示す図。

[図15]実施例1の特徴索引語抽出装置による「毛髪洗浄剤」に関する公開特許公報10件のマップ表示の具体例を示す図。

[図16]図15と同じ調査対象文書に関する、リスト出力の具体例を示す図。

[図17]実施例2の特徴索引語抽出装置によりTFIDF(S)を反映したマップの例を示す図。

[図18]実施例2の特徴索引語抽出装置によりTF(d)を反映したマップの例を示す図。

[図19]実施例2の特徴索引語抽出装置によるTFIDF平面図の例を示す図。

[図20]実施例2の特徴索引語抽出装置によるDF平面図の例を示す図。

[図21]実施例3の特徴索引語抽出装置により線形変換して出力したマップの例を示す図。

[図22]実施例3の特徴索引語抽出装置によりスケール変換して出力したマップの例を示す図。

[図23]実施例3の特徴索引語抽出装置により複合変換して出力したマップの例を示す図。

[図24]実施例3の特徴索引語抽出装置により複合変換して出力したマップの他の例を示す図。

[図25]実施例4の説明のために図10を書き直した図。

[図26]実施例4の例1における参照点の初期値を示す図。

[図27]実施例4の例1による変換で得たマップの例を示す図。

[図28]実施例4の例2における参照点の初期値を示す図。

[図29]実施例4の例2による変換で得たマップの例を示す図。

[図30]実施例4の例3における参照点の初期値を示す図。

[図31]実施例4の例3による変換で得たマップの例を示す図。

[図32]実施例4の例4による変換で得たマップの例を示す図。

[図33]実施例5の文書特徴分析装置のハードウェア構成を示す図。

[図34]実施例5の文書特徴分析装置の処理装置1の動作を示すフローチャート。

[図35]実施例5の文書特徴分析装置の出力装置4におけるマップ出力の動作を示すフローチャート。

[図36]実施例5の文書特徴分析装置によりある企業1社の文書特徴を示した図。

[図37]実施例5の文書特徴分析装置により同業界に属する3社の文書特徴を示した図。

符号の説明

[0034] 1:処理装置、2:入力装置、3:記録装置、4:出力装置、

120:索引語(d)抽出部、121:TF(d)演算部(索引語頻度算出手段)、142:IDF(P)演算部(第1、第3出現頻度算出手段)、150:類似度演算部、160:類似文書群S選出部、171:IDF(S)演算部(第2、第4出現頻度算出手段)、173:中心点算出部、180:特徴索引語抽出部、

a:独創語的着目語領域、b:専門語領域、c:類似文書群規定語領域、d:一般語領域

発明を実施するための最良の形態

[0035] 以下、本発明の実施の形態を、図面を参照して詳細に説明する。

[0036] <1. 語彙の説明等>

本明細書の中で使用する語彙を定義或いは説明する。

調査対象文書d:調査に係る、ある案件。例えば、特許第何号公報などの文書、或いはその集合。

比較対象文書群P:調査対象文書dと比較する対象の文書の集合。例えば、ある国及び期間における特許文書(公開特許公報など)のすべて、又はそこから無作為抽出された文書の集合である。ここでは調査対象文書dを含む場合について説明するが含んでいなくてもよい。

類似文書群S:調査対象文書dに類似な文書の集合。ここではdを含む場合について説明するが含んでいなくてもよい。また比較対象文書群Pの中から選出される場合について説明するが別の選出源文書群から選出されていても良い。

[0037] 図中の構成要素に付してある、d或いは(d)、P或いは(P)、又はS或いは(S)は、それぞれ調査対象文書、比較対象文書群、又は類似文書群の意味であり、以降判別しやすいように構成要素や動作にも付する。例えば、索引語(d)とは、調査対象文書dの索引語を意味する。

TF演算とはTerm Frequencyの計算のことであり、ある文書に含まれる索引語の当

該文書内の出現頻度(索引語頻度)の計算である。

DF演算とはDocument Frequencyの計算のことであり、ある文書に含まれる索引語で検索対象文書群から検索したときのヒット文書数(文書頻度)の計算である。

IDF演算とは、例えばDF演算結果の逆数、或いは逆数に検索対象文書群であるPないしSの文書数を乗じたものの対数などの計算である。

[0038] 以降の説明を簡素にするため、略号を決める。

d :調査対象文書。

p :比較対象文書群Pに属する文書。

N :比較対象文書群Pの全文書数。

N′ :類似文書群Sの文書の数。

TF(d):dの索引語による、dの中での出現頻度。

TF(P):pの索引語による、pの中での出現頻度。

DF(P):d又はpの索引語による、Pの中での文書頻度。

DF(S):dの索引語による、Sの中での文書頻度。

IDF(P):DF(P)の逆数×文書数の対数:ln[N/DF(P)]。

IDF(S):DF(S)の逆数×文書数の対数:ln[N′ /DF(S)]。

TFIDF:TFとIDFとの積。文書の索引語ごとに演算される。

類似度(類似率):調査対象文書dと、比較対象文書群Pに属する或る文書pとの類似の程度。

[0039] ここで、索引語とはいわゆるキーワードであり、文書の全部或いは一部から切り出される単語のことである。単語の切り出し方は従来から知られている方法や市販の形態素解析ソフトを活用して、助詞や接続詞を除き、意味ある品詞を抽出してもよいし、又索引語の辞書(シソーラス)のデータベースを事前に保持し該データベースから得られる索引語を利用してもよい。

また、対数としてここでは自然対数を用いたが、これに限らず常用対数等を用いてもよい。

[0040] <2. 索引語抽出装置の構成:図1、図2>

図1は本発明の一実施形態に係る特徴索引語抽出装置のハードウェア構成を示す

図である。

同図に示すように、本実施形態の特徴索引語抽出装置は、CPU(中央演算装置)およびメモリ(記録装置)などから構成される処理装置1、キーボード(手入力器具)などの入力手段である入力装置2、文書データや条件や処理装置1による作業結果などを格納する記録手段である記録装置3、および特徴索引語の抽出結果などをマップやリストなどで表示などする出力手段である出力装置4から構成される。

[0041] 図2は上記の特徴索引語抽出装置における構成と機能を詳細に説明する図である。

[0042] 処理装置1は、調査対象文書d読み出し部110、索引語(d)抽出部120、TF(d)演算部121、比較対象文書群P読み出し部130、索引語(P)抽出部140、TF(P)演算部141、IDF(P)演算部142、類似度演算部150、類似文書群S選出部160、索引語(S)抽出部170、IDF(S)演算部171、特徴索引語抽出部180などから構成される。

[0043] 入力装置2は、調査対象文書d条件入力部210、比較対象文書群P条件入力部220、抽出条件その他入力部230などから構成される。

[0044] 記録装置3は、条件記録部310、作業結果格納部320、文書格納部330などから構成される。文書格納部330は外部データベースや内部データベースを含んでいる。外部データベースとは、例えば特許庁でサービスしている特許電子図書館のIPDLや、株式会社パトリスでサービスしているPATOLISなどの文書データベースを意味する。又内部データベースとは、販売している例えば特許JP-ROMなどのデータを自前で格納したデータベース、文書を格納したFD(フレキシブルディスク)、CDROM(コンパクトディスク)、MO(光磁気ディスク)、DVD(デジタルビデオディスク)などの媒体から読み出す装置、紙などに出力された或いは手書きされた文書を読み込むOCR(光学的情報読み取り装置)などの装置及び読み込んだデータをテキストなどの電子データに変換する装置などを含んでいるものとする。

[0045] 出力装置4は、マップ作成条件読み出し部410、マップ用データ取り込み部412、リスト出力条件読み出し部420、リスト用データ取り込み部422、コメント追記条件読み出し部430、コメント追記部432、マップ・リスト・コメント複合出力部440などから構

成される。

[0046] 図1及び図2において、処理装置1、入力装置2、記録装置3、および出力装置4の間で信号やデータをやり取りする通信手段としては、USB(ユニバーサルシステムバス)ケーブルなどで直接接続してもよいし、LAN(ローカルエリヤネットワーク)などのネットワークを介して送受信してもよいし、文書を格納したFD、CDROM、MO、DVDなどの媒体を介してもよい。或いはこれらの一部、又はいくつかを組み合わせたものでもよい。

[0047] 次に、図2により本発明に係る一実施形態の特徴索引語抽出装置における機能を詳しく説明する。

[0048] <2−1. 入力装置2の詳細>

図2の入力装置2において、調査対象文書d条件入力部210は、入力画面などによって調査対象文書dの読み出しを行なう条件を設定する。比較対象文書群P条件入力部220は、入力画面などによって比較対象文書群Pの読み出しを行なう条件を設定する。抽出条件その他入力部230は、入力画面などによって調査対象文書d及び比較対象文書群Pの索引語抽出条件、TF演算の条件、IDF演算の条件、類似度演算の条件、類似文書の選出条件、マップ作成条件、リスト出力条件、コメント追記条件などを設定する。これら入力された条件は、記録装置3の条件記録部310へ送られ格納される。

[0049] <2−2. 処理装置1の詳細>

図2の処理装置1において、調査対象文書d読み出し部110は、条件記録部310の条件に基づいて、調査対象の文書を、文書格納部330より読み出す。読み出された調査対象文書dは、索引語(d)抽出部120に送られる。索引語(d)抽出部120は、条件記録部310の条件に基づいて、調査対象文書d読み出し部110で得られた文書から索引語の抽出を行ない、作業結果格納部320に格納する。

[0050] 比較対象文書群P読み出し部130は、比較対象となる複数の文書を、条件記録部310の条件に基づいて、文書格納部330より読み出す。読み出された比較対象文書群Pは、索引語(P)抽出部140に送られる。索引語(P)抽出部140は、条件記録部310の条件に基づいて、比較対象文書群P読み出し部130で得られた文書から索引

語の抽出を行ない、作業結果格納部320に格納する。

[0051] TF(d)演算部121は、条件記録部310の条件に基づいて、作業結果格納部320に格納された調査対象文書dについての索引語(d)抽出部120の作業結果を、TF演算する。得られたTF(d)のデータは、作業結果格納部320に格納され或いは直接類似度演算部150に送られる。

[0052] TF(P)演算部141は、条件記録部310の条件に基づいて、作業結果格納部320に格納された比較対象文書群Pについての索引語(P)抽出部140の作業結果を、TF演算する。得られたTF(P)のデータは、作業結果格納部320に格納され或いは直接類似度演算部150に送られる。

[0053] IDF(P)演算部142は、条件記録部310の条件に基づいて、作業結果格納部320に格納された比較対象文書群Pについての索引語(P)抽出部140の作業結果を、IDF演算する。得られたIDF(P)のデータは、作業結果格納部320に格納され、又は直接類似度演算部150に若しくは直接特徴索引語抽出部180に送られる。

[0054] 類似度演算部150は、条件記録部310の条件に基づいて、TF(d)演算部121、TF(P)演算部141、及びIDF(P)演算部142の演算結果を、それぞれから直接或いは作業結果格納部320から得て、比較対象文書群Pの文書それぞれの、調査対象文書dに対する類似度を演算する。得られた類似度は、比較対象文書群Pのそれぞれの文書に類似度データとして付され、作業結果格納部320或いは直接類似文書群S選出部160に送られる。

[0055] 類似度演算部150における類似度の演算は、各文書の索引語毎に、例えばTFIDF演算などの計算がなされ、比較対象文書群Pの文書それぞれの、調査対象文書dに対する類似度が計算される。TFIDF演算とは、TF演算結果とIDF演算結果の積である。類似度の演算方法は後で詳しく述べる。

[0056] 類似文書群S選出部160は、作業結果格納部320或いは直接類似度演算部150の結果から比較対象文書群Pの類似度演算結果を得て、条件記録部310の条件に基づいて類似文書群Sを選出する。類似文書群Sの選出は、例えば類似度の高い順に文書をソートし、条件に記された必要な数だけ選出する。選出された類似文書群Sは、作業結果格納部320或いは直接索引語(S)抽出部170に出力される。

[0057] 索引語(S)抽出部170は、作業結果格納部320或いは直接類似文書群S選出部160から類似文書群Sのデータ入力を得て、この類似文書群Sから、条件記録部310の条件に基づいて索引語(S)を抽出する。抽出された索引語(S)は、作業結果格納部320或いは直接IDF(S)演算部171に送られる。

[0058] IDF(S)演算部171は、作業結果格納部320或いは直接索引語(S)抽出部170から索引語(S)を得て、この索引語(S)を、条件記録部310の条件に基づいてIDF演算する。得られたIDF(S)は、作業結果格納部320に格納され或いは直接特徴索引語抽出部180に送られる。

[0059] 特徴索引語抽出部180は、条件記録部310の条件に基づいて、作業結果格納部320から、或いは直接IDF(S)演算部171の結果及びIDF(P)演算部142の結果から、条件に記された必要な数だけ、或いは条件に基づいた計算結果により選ばれた数だけ、索引語(d)を抽出する。ここで抽出された索引語を「特徴索引語」と称することにする。抽出された特徴索引語(d)は、作業結果格納部320に送られる。

[0060] ＜2−3. 記録装置3の詳細＞

図2の記録装置3において、条件記録部310は、入力装置2から得られた条件などの情報を記録し、処理装置1或いは出力装置4の要求に基づき、それぞれに必要なデータを送る。作業結果格納部320は、処理装置1における各構成要素の作業結果を格納し、処理装置1の要求に基づき、必要なデータを送る。

[0061] 文書格納部330は、入力装置2或いは処理装置1の要求に基づき、外部データベース或いは内部データベースから得た、必要な文書データを格納し、提供する。

[0062] ＜2−4. 出力装置4の詳細＞

図2の出力装置4において、マップ作成条件読み出し部410は、条件記録部310の条件に基づいて、マップの作成条件を読み出し、マップ用データ取り込み部412に送る。リスト出力条件読み出し部420は、条件記録部310の条件に基づいて、リストの出力条件を読み出し、リスト用データ取り込み部422に送る。コメント追記条件読み出し部430は、条件記録部310の条件に基づいて、コメントの追記条件を読み出し、コメント追記部432に送る。

[0063] マップ用データ取り込み部412は、マップ作成条件読み出し部410の条件に従い、

作業結果格納部320より、特徴索引語抽出部180の作業結果を取り込む。取り込まれた特徴索引語データは、作業結果格納部320或いは直接マップ・リスト・コメント複合出力部440に送られる。

[0064] リスト用データ取り込み部422は、リスト出力条件読み出し部420の条件に従い、作業結果格納部320より、特徴索引語抽出部180の作業結果を取り込む。取り込まれたリスト用データは、作業結果格納部320或いは直接マップ・リスト・コメント複合出力部440に送られる。

[0065] コメント追記部432は、コメント追記条件読み出し部430の条件に従い、キーボードやOCRなどの外部入力装置から直接、或いは文書格納部330の内部データベースに事前に用意した、調査対象文書dに対する評価のコメントとして追記するためのデータを準備する。準備されたコメント用データは、作業結果格納部320或いは直接マップ・リスト・コメント複合出力部440に送られる。

[0066] マップ・リスト・コメント複合出力部440は、マップ用データ取り込み部412から出力される条件とデータ、リスト用データ取り込み部422から出力される条件とデータ、及びコメント追記部432から出力される条件とデータをそれぞれ直接或いは作業結果格納部320より得て、マップ・リスト・コメントを複合的に出力する場を作る。同時に、特徴索引語抽出部180の作業結果を、マップ上に表示し、一覧リストに出力し、及びコメント或いはそれらの一部を表示、印刷、若しくはデータで格納できるように出力する。

[0067] マップ・リスト・コメント複合出力部440において出力するマップの特徴的な一例は、特徴索引語抽出部180において抽出された調査対象文書dの特徴索引語の各々について、比較対象文書群PにおけるIDF(P)演算部142の演算結果を横軸の値とし、調査対象文書dに類似な類似文書群SにおけるIDF(S)演算部171の演算結果を縦軸の値として、二次元のIDF(P)−IDF(S)平面(以下、IDF平面と呼ぶ)上に分布させたマップである。図11以降の説明において詳細を後述する。該IDF平面上で表わされた特徴索引語の分布状況から、調査対象文書dの性格を読み取ることができる。

[0068] ＜3. 索引語抽出装置の動作＞

図3、図4、及び図5は上記の特徴索引語抽出装置における動作を説明する図である。

[0069] ＜3−1．入力動作：図3＞

図3は、入力装置2における条件設定の動作手順を示すフローチャートである。後述の図6ー図9に、入力装置により入力する条件設定の操作画面が例示されている。まず初期化（ステップS201）のあと、入力する条件を区別する（ステップS202）。オペレータが調査対象文書dの条件入力を選定したときは、調査対象文書d条件入力部210において調査対象文書dの条件入力を受けつける（ステップS210）。次に、入力された条件が図6のような表示画面でオペレータにより確認され、よければ画面上の「設定」が選ばれるので、入力された条件を条件記録部310で格納し（ステップS310）、悪ければ「戻る」が選ばれるので、ステップS210に戻る（ステップS211）。

[0070] 一方ステップS202においてオペレータが比較対象文書群Pの条件入力を選定したときは、比較対象文書群P条件入力部220において比較対象文書群Pの条件入力を受けつける（ステップS220）。次に、入力された条件が図7のような表示画面でオペレータにより確認され、よければ画面上の「設定」が選ばれるので、入力された条件を条件記録部310で格納し（ステップS310）、悪ければ「戻る」が選ばれるので、ステップS220に戻る（ステップS221）。

[0071] 又、ステップS202においてオペレータが抽出条件その他の入力を選定したときは、抽出条件その他入力部230において抽出条件その他の入力を受けつける（ステップS230）。次に、入力された条件が図8や図9のような表示画面でオペレータにより確認され、よければ画面上の「設定」が選ばれるので、入力された条件を条件記録部310で格納し（ステップS310）、悪ければ「戻る」が選ばれるので、ステップS230に戻る（ステップS231）。該ステップS230においては、索引語（d）の抽出条件及び類似文書群Sの選出条件と、特徴索引語等の出力条件との両方を設定する。

[0072] ＜3−2．特徴索引語の抽出動作：図4＞

図4は、処理装置1の動作を示すフローチャートである。まず初期化（ステップS101）のあと、条件記録部310の条件に基づいて、文書格納部330から読み出す文書を、調査対象文書dと比較対象文書群Pに区別する（ステップS102）。読み出す文書が

調査対象文書dであるとき、調査対象文書d読み出し部110において調査対象文書を文書格納部330より読み出す(ステップS110)。次に、索引語(d)抽出部120において調査対象文書dの索引語抽出を行なう(ステップS120)。引き続き、抽出された索引語の各々について、TF(d)演算部121においてTF演算をする(ステップS121)。

[0073] 一方ステップS102で、読み出す文書が比較対象文書群Pであるとき、比較対象文書群P読み出し部130において比較対象文書群Pを読み出す(ステップS130)。次に、索引語(P)抽出部140において比較対象文書群Pの索引語抽出を行なう(ステップS140)。引き続き、抽出された索引語の各々について、TF(P)演算部141においてTF演算をする(ステップS141)とともに、IDF(P)演算部142においてIDF演算をする(ステップS142)。

[0074] 次に、TF(d)演算部121の出力のTF(d)演算結果と、TF(P)演算部141の出力のTF(P)演算結果、及びIDF(P)演算部142の出力のIDF(P)演算結果を基に、類似度演算部150により、類似度の演算を行なう(ステップS150)。この類似度の演算は、入力装置2から入力された条件に基づき、類似度算出のための類似度算出モジュールを外部記録部310から呼び出してきて実行する。

[0075] 類似度演算の具体的な一例を説明すると以下の通りである。今、dを調査対象文書とし、pを比較対象文書群Pの個々の文書とする。これら文書d及びpに対する演算の結果、文書dから切り出された索引語を「赤」「青」「黄」とする。また、文書pから切り出された索引語を「赤」「白」とする。その場合、文書d中の索引語の索引語頻度をTF(d)とし、文書p中の索引語の索引語頻度をTF(P)とし、比較対象文書群Pから得た索引語の文書頻度をDF(P)とし、全文書数を50とする。このとき、例えば、

[0076] [表1]

索引語及びTF(d)	赤(1), 青(2), 黄(4)
索引語及びTF(P)	赤(2), 白(1)
索引語及びDF(P)	赤(30), 青(20), 黄(45), 白(13)

[0077] であるとする。TF＊IDF(P)を各文書の索引語毎に計算して、ベクトル表現を算出

する。この結果は文書ベクトルd及びpについて、

[0078] [表2]

	赤	青	黄	白
d	(1* ln(50/30))	(2* ln(50/20))	(4* ln(50/45))	0
p	(2* ln(50/30))	0	0	(1* ln(50/13))

[0079] となる。このベクトルd及びp間の余弦(又は距離)の関数を取れば、文書ベクトルd及びp間の類似度(又は非類似度)が得られる。なお、ベクトル間の余弦(類似度)は値が大きいほど類似度合いが高いことを意味し、ベクトル間の距離(非類似度)は値が小さいほど類似度合いが高いことを意味する。得られた類似度は、作業結果格納部320に格納されるとともに、類似文書群S選出部160に送られる。

[0080] 次に、類似文書群S選出部160により、ステップS150にて類似度演算した文書を類似度の順に並べ替え、抽出条件その他入力部230において設定した条件に沿った数の類似文書群Sを選出する(ステップS160)。

[0081] 次に、類似文書群Sの索引語(S)抽出部170により、ステップS160にて選出した類似文書群Sの索引語(S)を抽出する(ステップS170)。

[0082] 次に、索引語(d)の各々について、IDF(S)演算部171により、類似文書群SにおけるIDF演算をする(ステップS171)。

[0083] 次に、ステップS171によるIDF(S)演算の結果と、ステップS142によるIDF(P)演算の結果とから、特徴索引語を抽出する(ステップS180)。

[0084] <3−3. 出力動作:図5>

図5は、出力装置4による、マップ、リスト、及びコメントの出力の動作手順を示すフローチャートである。まず初期化(ステップS401)のあと、条件記録部310から、マップ作成条件と、リスト出力条件と、コメント追記条件に区別して条件の読み出しを開始する(ステップS402)。

[0085] 出力装置のマップ作成条件読み出し部410で条件記録部310からマップ作成条件を読み出したとき(ステップS410)、マップを必要とする条件であったら(ステップS411)、作業結果格納部320からマップ用データ取り込み部412へのマップ用データ

取り込みを行なう(ステップS412)。次に、マップ作成条件読み出し部410のマップ作成条件に沿って、マップを作成し(ステップS413)、マップ・リスト・コメント複合出力部440に送る。

[0086] 一方、出力装置のリスト出力条件読み出し部420で条件記録部310からリスト出力条件を読み出したとき(ステップS420)、リストを必要とする条件であったら(ステップS421)、作業結果格納部320からリスト用データ取り込み部422によりリスト用データ取り込みを行なう(ステップS422)。次に、リスト出力条件読み出し部420のリスト出力条件に沿って、リストを生成し(ステップS423)、続いて、マップ・リスト・コメント複合出力部440に送る。

[0087] また一方、出力装置のコメント追記条件読み出し部430で条件記録部310からコメント追記条件を読み出したとき(ステップS430)、コメントを必要とする条件であったら(ステップS431)、マップ・リスト・コメント複合出力部440にてコメントを追記できる枠を準備し、該枠内に、キーボードから或いはOCRから、手入力するか、或いは文書格納部330の内部データベースにある事前に準備された定型文データを使って、コメント追記を生成し(ステップS433)、続いて、マップ・リスト・コメント複合出力部440に送る。

[0088] ステップS411でマップを表示する条件でなかったら、又はステップS421でリストを出力する条件でなかったら、又はステップS431でコメントを追記する条件でなかったら、それぞれその時点で終了し、マップ・リスト・コメント複合出力部440へはデータを送らない。

[0089] <3−4. 入力画面:図6ー図9>

図6は、調査対象文書dの入力条件設定画面の表示例を示す図である。

[0090] 図6においては、「対象文書」のウインドの「調査対象文書」と「比較対象文書群」の中から「調査対象文書」を選び、次に「文書種別」のウインドの「公開特許」「登録特許」「実用新案」「学術文献」などの中から「公開特許」を選び、次に「データの読み出し」のウインドの「自社DB1」「自社DB2」「特許庁IPDL」「PATOLIS」「他商用DB1」「他商用DB2」「FD」「CD」「MO」「DVD」「その他」などの中から「FD」を選び、更に「FD」の「文書1」「文書2」「文書3」「文書4」「文書5」「文書6」などの中から「文書3」を

選んだ状態の例が示されている。この例のような入力条件設定画面における設定条件が、調査対象文書d条件入力部210で入力される。

[0091] 図7は、比較対象文書群Pの入力条件設定画面の表示例を示す図である。図7においては、「対象文書」のウインドの「調査対象文書」と「比較対象文書群」などの中から「比較対象文書群」を選び、次に「文書種別」のウインドの「公開特許」「登録特許」「実用新案」「学術文献」などの中から「公開特許」と「登録特許」の両方を選び、次に「抽出内容」のウインドの「請求項」「従来技術」「発明の課題」「手段・効果」「実施例」「図の説明」「図面」「要約」「書誌事項」「経過情報」「登録情報」「その他」などの中から「請求項」と「要約」の両方を選び、次に「データの読み出し」のウインドで前述と同じ項目の中から「自社DB1」を選んだ状態の例が示されている。この例のような入力条件設定画面における設定条件が、比較対象文書群P条件入力部220で入力される。

[0092] 図8は、索引語抽出条件および類似文書群選出条件の設定画面の表示例を示す図である。図8においては、「索引語抽出条件」のウインドの「自社キーワード切出1」「自社キーワード切出2」「商用キーワード切出1」「商用キーワード切出2」などの中から「自社キーワード切出1」を選び、次に「類似度算出方法」のウインドの「類似度1」「類似度2」「類似度3」「類似度4」「類似度5」「類似度6」などの中から「類似度1」を選び、次に「類似文書群選出」のウインドの「類似文書数」「非類似文書数」などの中から「類似文書数」を選び、更に「上位100件」「上位1000件」「上位3000件」「上位5000件」「数値入力」などの中から「上位3000件」を選んだ状態の例が示されている。この例のような抽出条件設定画面における設定条件が、抽出条件その他入力部230で入力される。

[0093] 図9は、特徴索引語抽出装置の出力条件設定画面の表示例を示す図である。図9においては、「マップ算出方法」のウインドの「X軸」に「X軸:比較対象文書群IDF」及び「Y軸」に「Y軸:類似文書群IDF」を選び、次に「マップ形式」のウインドの「マップ1枚」「マップ2枚」「マップ1枚・リスト付」「マップ2枚・リスト付」「マップ1枚・コメント付」「マップ2枚・コメント付」「マップ1・リスト・コメント付」「マップ2・リスト・コメント付」などの中から「マップ1枚」を選び、次に「出力データ」のウインドの「独創的着目語」「専門語」「類似文書群規定語」などの中から「独創的着目語」を選び、更に「なし」「上位5個」

「上位10個」「上位15個」「上位20個」「数値入力」などの中から「上位20個」を選んだ状態の例が示されている。次に「コメント」のウインドの枠内の「(自由記入)」には無記入にした。こうして抽出条件その他入力部230より、出力条件が入力される。

[0094] ＜4. 実施例1＞

＜4−1. マップの性質:図10＞

図10は、実施例1の索引語抽出装置により出力したマップの性質を説明するための概念図である。このマップは、調査対象文書dの索引語(d)のうち特徴索引語抽出部180で抽出された索引語(以下、特徴索引語という)を、マップ・リスト・コメント複合出力部440で出力し、表示手段により表現するものである。マップは、特徴索引語の各々について、それぞれ、横軸の値に比較対象文書群PにおけるIDF(P)演算部142の演算結果を、縦軸の値に類似文書群SにおけるIDF(S)演算部171の演算結果を取って、IDF平面上に配置したものである。

[0095] 図10を説明する。図10において、X−Y平面は、X軸がIDF(P)の値で、Y軸がIDF(S)の値で作る平面である。比較対象文書群Pの文書数をN、類似文書群Sの文書数をN′とすれば、IDF(P)の最大値$\beta_1 = \ln N$、IDF(S)の最大値$\beta_2 = \ln N'$である。

平面の原点をDとする。Y＝Xの直線と、Y＝β_2の線との交点をAとする。Y＝β_2の線と、X＝β_1の線の交点をBとする。Y−β_2＝X−β_1の直線がX軸を切る点をCとする。従って、四角形ABCDは、平行四辺形である。$\alpha = \beta_1 - \beta_2 = \ln(N/N')$とすると、平行四辺形ABCDの各頂点の値は、それぞれ、D＝(0, 0)、B＝(β_1, β_2)、A＝(β_2, β_2)、C＝(α, 0)である。

[0096] 線分ABは、Y＝β_2、線分ADは、Y＝Xの直線である。線分BCは、Y−β_2＝X−β_1の直線である。線分DCは、Y＝0の直線である。

[0097] 図10において、X座標はIDF(P)の値であり、Xの値が0付近すなわちD付近は、比較対象文書群Pのほとんどに存在する索引語が配置される領域である。X座標が$\beta_1 = \ln N$の内側は、比較対象文書群Pにもほとんど存在しない索引語の領域で、X座標が$\alpha = \ln(N/N')$の内側すなわちCの内側は、比較対象文書群Pにも類似文

書群Sの文書数N′ 相当の数が存在する索引語の領域である。一方、Y座標はIDF(S)の値であり、Yの値が0付近すなわちD付近は、類似文書群Sのほとんどに存在する索引語の領域である。Y座標が$\beta_2=\ln N'$ の線分ABの内側は、類似文書群Sの中にはほとんど存在せず、ほぼ調査対象文書dにしか存在しない索引語の領域である。

[0098] 図10において、比較対象文書群Pにおける文書頻度DF(P)が小さい、即ち珍しい索引語は、IDF(P)が大きいため、図10上の右側に現れる。DF(P)が大きい、即ち頻繁に用いられる索引語は、IDF(P)が小さいため、図10上のY軸の近くに現れる。従って、比較対象文書群Pにおいて珍しい索引語ほど右に現れ、比較対象文書群Pにおいて頻繁に用いられる索引語ほど左に現れる。二次元平面上では類似文書群Sが比較対象文書群Pの部分集合であることによる制限が課せられるため、図10の右側では線分BCで切られる領域内部にしか索引語の点は存在しない。

[0099] 同様に、類似文書群Sにおける文書頻度DF(S)が1件しかない索引語、即ち調査対象文書d自身にしか含まれていない索引語は、IDF(S)が大きいため、図10上のBA線上に現れる。DF(S)が1より大きいと、索引語はBA線より下に位置する。逆に、類似文書群Sの全ての文書に存在する索引語は、IDF(S)＝0 のため、図10上のDC線上、すなわちy＝0 の線上に現れる。従って、Sにおいて珍しい索引語ほど上に現れ、Sにおいて頻繁に用いられる索引語ほど下に現れる。

[0100] ここで線分BCは次により導出される。類似文書群Sが比較対象文書群Pの部分集合であることより、

$DF(P) \geqq DF(S)$

である。また、IDFの上記定義より、

$DF(P)=N\ \exp[-IDF(P)]$、

$DF(S)=N'\ \exp[-IDF(S)]$

である。これらの関係式より、境界線の式としてy＝x－α、即ち$y-\beta_2=x-\beta_1$ が得られる。

[0101] 類似文書群Sの文書数に依存せず、一様に含まれる索引語の場合、その索引語は図10の線分DA(直線Y＝X)上に現れる。ここで一様とは、計測対象とする文書群Q

の文書数N_Qを変化させる時、

$DF(Q)=N_Q/k$（kは1より大なる定数）

が成立するQを一様又は空間一様性のある文書群、また、その様な性質を持つ索引語を、空間一様性を持つ索引語と呼ぶ。Q＝P，Sに対して一様性を仮定すると、

$\ln k=\ln[N/DF(P)]=\ln[N'/DF(S)]$

より、直線Y＝Xが得られる。

実際には、多くの索引語は類似文書群Sよりも膨大な文書群である比較対象文書群Pにおいても頻出するから、線分DAの下方領域に出現するのが普通であり、特異なものだけがこの線分の上側に浮かび上がることになる。このうち特に、図10内の線分BAの半分位の高さより上側の領域にあっては、比較対象文書群Pにおいては珍しくないが、類似文書群Sにおいては珍しい索引語が出現する。この傾向によりA付近の領域は独創的着目語領域と言ってよい。

[0102] 図10において、線分AD左方の充分外側の領域にも索引語の点は存在可能であるが、次のことを考え合わせると、該領域を索引語の点の非存在領域として扱っても、調査対象文書dの性質解読に支障を来たすものではない。すなわち、該領域は、独創的着目語領域Aの遠方の領域なので、もし出現したとしても、かなり特異な索引語であること、y軸近傍には$DF(S)\geqq DF(P)-N+N'$の制限から導かれる存在限界線：

$Y=-\ln(\gamma\exp(-x)-\gamma+1)$、但し$\gamma=N/N'$

があり、同線に近いこと、観測的事実として、類似文書群Sの類似度が十分高い場合には該領域には索引語が観測されなかったことなどをあわせて、事実上、非存在領域と帰結される。

[0103] 以上のように、調査対象文書dから抽出された特徴索引語は、図10のIDF平面の右に行くほど比較対象文書群Pでの文書頻度は低く、上に行くほど類似文書群Sでの文書頻度が低い。そこで、図10における各領域には、次のような性質を持つ索引語が配置されるため、該IDF平面上の点の分布状況から、調査対象文書dの、比較対象文書群Pの中に対する位置付けや性格を読み取ることができる。

[0104] 専門語領域b：比較対象文書群Pにおいても類似文書群Sにおいても使用頻度の

低い索引語が現れる領域。すなわち調査対象文書dに含まれる高度に専門的な内容、又はこれに直結する概念を記述する索引語の出現する領域。本発明の第1エリアに含まれる。

[0105] 独創的着目語領域a:比較対象文書群Pにおける出現頻度の高さの割には、類似分野ではあまり着目されていなかった概念を示す索引語の出現する領域。本発明の第2エリアに含まれる。

[0106] 類似文書群規定語領域c:類似文書群Sでほとんどの文書が持ち、従って比較対象文書群Pにおいてもそれに相当する数の文書が持っている、類似文書群Sの性質を表わすのに極めて自然な索引語が現れる領域。例えば技術文書を調査対象とした場合であれば、この類似文書群規定語を見れば、類似文書群S及び調査対象文書dの技術分野を知ることができる。本発明の第3エリアに含まれる。

[0107] 一般語領域d:比較対象文書群Pと類似文書群Sの両方において頻出する索引語が現れる領域。比較対象文書群Pとの比較において調査対象文書dの性格を分析する際には、重要度が低いことが多い。

[0108] <4−2. マップ出力例1:図11(外部補助記憶装置)>

図11は、実施例1の特徴索引語抽出装置において、調査対象文書dとして「外部補助記憶装置」に関する公開特許公報を1件選んだときの、マップ表示の具体例である。このマップは本発明の性格表現図に相当する(以下のマップも同様)。比較対象文書群Pとして、過去10年間の特許公報及び公開特許公報約464万件を選び、抽出内容には特許請求の範囲と要約を選び、索引語抽出は自社キーワード切り出し1(市販の索引語切出しツール)を選び、類似度算出方法には、文書ベクトルの成分ごとにTFIDFを計算し調査対象文書dと比較対象文書群Pのそれぞれとの余弦を計算する方法を選び、類似文書群S選出には類似度の上位3000件を選び、マップ算出方法にはX軸:比較対象文書群Pに対するIDFと、Y軸:類似文書群Sに対するIDFを選び、マップ出力位置にマップ1枚を選んだ結果、表示されたものである。

[0109] 図11から、図10にて示した独創的着目語領域aには、「絵」「ホログラム」「欲求」「プラスチック」「外面」などの特徴索引語を見つけ、同じく専門語領域bには、該当する特徴索引語を見つけることができず、又同じく類似文書群規定語領域cには、「コンテ

ンツ」「編集」などの特徴索引語を見つけることができる。

[0110] ＜4−3. リスト出力例1：図12（外部補助記憶装置）＞

図12は、図11と同じ調査対象文書に関する、リスト出力の具体例である。このリストは本発明の性格表現図に相当する（以下のリストも同様）。

[0111] 各領域において出力すべき索引語は、例えば次のように求められる。

各領域に応じて変換M：(X, Y)→(X′, Y′)が与えられる時、

(s/100) Exp[Y′]＜2

なる点を、X′で降順に抽出する。但し、

(p/100) Exp[X′]≧2

なる点に限る。

[0112] 各領域から抽出するための上記変換M(X′, Y′)は次で与えられる：

独創的着目語領域a……(X, X−Y)、

専門語領域b……(Y, Y−X＋α)、

類似文書群規定語領域c…(X, Y)、

一般語領域d……(Y−X＋α, Y)。

但し、α＝ln(N／N′)。

[0113] 例えば類似文書群規定語を抽出する場合は、比較対象文書群Pにおける文書数Nに対する文書頻度DF(P)の割合がp／2(%)以下で、且つ類似文書群Sにおける文書数N′に対する文書頻度DF(S)の割合が、s／2(%)を超える索引語が抽出されることになる。図12では、p＝s＝25として索引語を抽出した。

独創的着目語、専門語及び一般語に対する変換値(X′, Y′)はそれぞれ類似文書群規定語領域c付近に写像したものであるので、同様の抽出条件を用いることにより各領域の索引語が抽出される。

[0114] なお、抽出条件は上記に限らず、例えば、

$PDF(w_i, P) = (p/100)Exp[X'] - 1$、

$PDF(w_i, S) = (s/100)Exp[Y'] - 1$

とおいて、

$PDF(w_i, P) \geqq 1$のとき、

$X'' = \ln PDF(w_i, P)$、

$0 < PDF(w_i, P) < 1$のとき、

$X'' = -1$、

$PDF(w_i, P) \leqq 0$のとき、

$X'' = -2$

のように離散化し(Y′ についても同様)、$Y'' < 0$且つ$X'' \geqq 0$なる索引語を、X'' 値の降順に抽出しても同様の結果を得ることができる。

[0115] 図12で出力されたデータを調べると、図10にて示した独創的着目語領域aには、「絵」「ホログラム」「制作」「プラスチック」「外面」などの特徴索引語が含まれ、同じく専門語領域bには、該当する特徴索引語がなく、又同じく類似文書群規定語領域cには、「コンテンツ」「編集」などの特徴索引語が含まれていることがわかる。

[0116] 図11或いは図12から、本発明の特徴索引語抽出装置において、調査対象文書dの「外部補助記憶装置」に関する公開特許公報にとって特徴のある索引語を調べた結果、「プラスチック」「外面」「ホログラム」「絵」などが独創的概念着目語であり、専門語はなく、「コンテンツ」「編集」などが類似文書群規定語であることが分かる。

なお、出力する索引語の個数は、各領域についてそれぞれ複数であることが望ましいが、単数でも良いし、本出力例のように該当する索引語がない領域については0でもよい。

[0117] <4−4. マップ出力例2:図13(緊急通報)>

図13は、図11の条件と同じで、調査対象文書dとして「緊急通報」に関する公開特許公報を1件選んだときの、マップ表示の具体例である。

[0118] 図13から、独創的着目語領域aには、「既知」「デファレンシャル」「老齢」「基準局」「DGPS」などの特徴索引語を見つけ、専門語領域bには、点Bから若干離れたところに「消防署」などの特徴索引語を見つけ、類似文書群規定語領域cには、「通報」「緊急」「事態」などの特徴索引語を見つけることができる。

[0119] <4−5. リスト出力例2:図14(緊急通報)>

図14は、図13と同じ調査対象文書に関する、リスト出力の具体例である。図14で出力されたデータを調べると、独創的着目語領域aには、「デファレンシャル」「既知」

「手順」などの特徴索引語が含まれ、専門語領域bには、「消防署」などの特徴索引語を見つけ、類似文書群規定語領域cには、「事態」「通報」「緊急」「センタ」「電話機」などの特徴索引語が含まれていることがわかる。

[0120] 図13或いは図14から、本発明の特徴索引語抽出装置において、調査対象文書dの「緊急通報」に関する公開特許公報にとっては、「デファレンシャル」「既知」などが独創的着目語であり、「消防署」が専門語であり、「事態」「通報」「緊急」などが類似文書群規定語である特徴索引語であることが分かる。

[0121] ＜4−6. マップ出力例3:図15(毛髪洗浄剤)＞

図15は、図11の条件と同じで、調査対象文書dとして「毛髪洗浄剤」に関する公開特許公報を10件選んだときの、マップ表示の具体例である。

[0122] 図15から、独創的着目語領域aには、「高齢」「クシ」「行為」「ml」「カリ」「工程」「滞留」「ブラシ」などの特徴索引語を見つけ、専門語領域bには、「フライアウェイ」「ジアリルアンモニウム」「メタクリロイルエチル」「ポリオキシエチレンジオレイン」などの特徴索引語を見つけ、類似文書群規定語領域cには、「両性」「毛髪」「アニオン」「アルケニル」「脂肪酸」などの特徴索引語を見つけることができる。

[0123] ＜4−7. リスト出力例3:図16(毛髪洗浄剤)＞

図16は、図15と同じ調査対象文書に関するリスト出力の具体例である。図16で出力されたデータを調べると、独創的着目語領域aには、「クシ」「ml」「カリ」「薬効」「高齢」「行為」「外用」などの特徴索引語が含まれ、専門語領域bには、「フライアウェイ」「ポリオキシエチレンジオレイン」「メチルカルボキシベタイン」「ジアリルアンモニウム」などの特徴索引語が含まれ、類似文書群規定語領域cには、「両性」「毛髪」「ヒドロキシアルキル」「泡」「皮膚」「アニオン」「カチオン」「脂肪酸」などの特徴索引語が含まれていることがわかる。

[0124] 図15或いは図16から、本発明の特徴索引語抽出装置において、調査対象文書dの「毛髪洗浄剤」に関する公開特許公報にとっては、「高齢」「クシ」が独創的着目語であり、「フライアウェイ」「ポリオキシエチレンジオレイン」は専門語であり、「両性」「毛髪」が類似文書群規定語であることが分かる。

[0125] こうして、本発明の特徴索引語抽出装置を利用すれば、人間が調査対象文書の内

容を読むことなく、その文書の性格を的確に表す特許マップを提供することができる。

[0126] ＜4−8. コメント出力＞

本発明の特徴索引語抽出装置による出力は、上記のマップやリストに限らず、代表的な索引語を用いて調査対象文書dの性格を解説するコメント文を自動生成して出力しても良い。コメント文は、例えば、図12，図14，図16で出力したリストの上位数個の索引語を用いて、「**、**（類似文書群規定語領域cの索引語）に関する技術分野において、**、**（専門語領域bの索引語）に関わる専門的な概念・技術を利用し、**、**（独創的着目語領域aの索引語）の観点に着目した文書」のように生成する。

また例えば専門語領域bに索引語が現れなかったときは、コメント文は専門語に関する記述を除き、「**、**（領域cの索引語）に関する技術分野において、**、**（領域aの索引語）の観点に着目した文書」のように生成する。

また例えば独創的着目語領域aに索引語が現れなかったときは、コメント文は独創的着目語に関する記述を除き、「**、**（領域cの索引語）に関する技術分野において、**、**（領域bの索引語）に関わる専門的な概念・技術を利用した文書」のように生成する。

また例えば独創的着目語領域a、専門語領域bいずれも索引語が現れなかったときは、コメント文は独創的着目語及び専門語に関する記述を除き、「**、**（領域cの索引語）に関する技術分野に属する文書」のように生成する。

[0127] このコメント文は、上記のマップや表と一緒に出力しても良いし、コメントのみを出力しても良い。また、出力する索引語の個数は、各領域についてそれぞれ複数であることが望ましいが、単数でも良いし、該当する索引語がない領域については0でもよい。

[0128] ＜5. 実施例2＞

図17ー図20は、実施例2の特徴索引語抽出装置により出力したマップの例を示す図である。特徴索引語抽出装置の具体的な構成は実施例1と同様であるので詳細な説明を省略し、主な相違点について説明する。

[0129] ＜5−1．TF又はTFIDF重み付け：図17、図18＞

図11に示したIDF平面図で、抽出された特徴索引語を単純にマップ表示しても、調査対象文書dにおいてどの索引語が重視されているのかは不明である。そこで、調査対象文書d中における当該特徴索引語の出現頻度TF(d)或いはこれとIDF(S)との積であるTFIDF(S)を索引語の位置づけデータに反映させる。反映のさせ方としては、当該特徴索引語のマップ上の存在点でサイズ(表示の大きさ)を変えたり、表示の形を変えたり、或いは色を変えて表示して、重視される特徴索引語の視覚化を図る。反映のさせ方としては他にも、各索引語の出現頻度TF(d)又はTFIDF(S)をZ成分とし、3次元グラフィックにより3次元座標を表示する方法などが考えられる。

[0130] この場合には、マップ作成条件の一つとして、異なる特徴索引語に対し、出現頻度順にサイズや形や色を自動的に割り当てる情報を条件記録部310に格納しておけばよい。マップ表示の時、入力装置からの指示により、特徴索引語抽出部180に読み出してきて、特徴索引語抽出部180においてその割り当て処理を行って出力させることが出来る。このマップ出力信号は、TF(d)またはTFIDF(S)を反映した出現頻度反映信号である。

[0131] 図11に示された特徴索引語に対して、このような処理を行った例を図17及び図18にそれぞれ示す。図17は、TFIDFが上位20までの特徴索引語について、○印を付して表示した例を示す図である。図18は、TF値が上位10までの特徴索引語について、サイズの大きい◇印を付して表示した例を示す図である。

[0132] ＜5−2．TFIDF及びDF平面図：図19、図20＞

図19と図20は、図11と同じく調査対象文書dとして「外部補助記憶装置」に関する公開特許公報を1件選ぶとともに、文書群における各索引語の出現頻度の関数値の取り方を実施例1とは変えて出力したものである。

[0133] 図19は、調査対象文書dの索引語(d)の各々について、横軸に比較対象文書群Pに対するTFIDF(TF(d)とIDF(P)との積)をとり、縦軸に類似文書群Sに対するTFIDF(TF(d)とIDF(S)との積)をとって、分布させたもの(以下TFIDF平面図という)である。

[0134] 図19によってTF(d)を加味して評価すれば、「データ」「コンテンツ」「編集」などを

類似文書群規定語と評価でき、「物」「算出」「適合」「IC」「プラスチック」などを独創的着目語と評価することができる。しかし、原点付近にほとんどの点が集中するため、点の分布状況から直接且つ容易に調査対象文書dの性質を論じることは困難である。実施例1の図11などの表示をこの図19と対比すると明らかなように、実施例1によるIDF平面図の方が、調査対象文書dの性質の容易で直接的な解読には好ましい、ということがわかる。原点付近への点の集中を回避するための一法としては、TFIDFの対数をとって座標上に配置することも考えられる。

[0135] 図20は、調査対象文書dの索引語(d)の各々について、横軸に比較対象文書群PにおけるDFを文書数Nで除したものをとり、縦軸に類似文書群SにおけるDFを文書数N′で除したものをとって、分布させたもの(以下DF平面図という)である。図20によってDFに基づき評価すれば、「データ」「記憶」「情報」「媒体」「編集」「コンテンツ」などを類似文書群規定語と評価でき、「物」「内部」「プラスチック」などを独創的着目語と評価することができる。しかし、この場合も、原点付近にほとんどの点が集中するため、点の分布状況から直接且つ容易に調査対象文書dの性質を論じることは困難である。実施例1の図11などの表示をこの図20と対比すると明らかなように、実施例1によりDF値を逆冪の対数で変換したIDF平面図の方が、調査対象文書dの性質の容易で直接的な解読には好ましい、ということがわかる。原点付近への点の集中を回避するための一法としては、DFそのものの対数をとって座標上に配置することも考えられる。

[0136] 文書群における索引語の出現頻度は、上記DFに限らず、例えば検索対象文書群から検索したときに当該索引語がヒットした延べ回数を用いてもよい。

[0137] <6. 実施例3:図の変形>

図21乃至図24は、実施例3の特徴索引語抽出装置により出力したマップの例を示す図である。特徴索引語抽出装置の具体的な構成は実施例1と同様であるので詳細な説明を省略し、主な相違点について説明する。

[0138] 上述の実施例1又は2により、調査対象文書を評価する者は、特徴索引語抽出装置の出力結果を観察すれば、該文書の内容を読むことなく、該文書の大きな傾向と

しての性格を読取ることができる。

[0139] 但し、観察者が不慣れの場合には、図11、図13及び図15（以下代表して図11のみ示すことがある）等のように、境界線BC等がX軸に対して斜交していると、領域を特定しにくい場合がある。特に類似文書群Sが比較対象文書群Pの部分集合である場合、例えばある索引語を比較対象文書群Pで検索したときのヒット文書数DF（P）は、同じ索引語を類似文書群Sで検索したときのヒット文書数DF（S）より小さい数にはなり得ない。また、ある索引語を比較対象文書群Pで検索したときにヒットしない文書数N−DF（P）は、同じ索引語を類似文書群Sで検索したときにヒットしない文書数N′−DF（S）より小さい数にはなり得ない。従って例えば上記DF（P）を直交座標のX軸に、上記DF（S）をY軸にとろうとすると、X≧YかつN−X≧N′−Yの領域にのみ各索引語が配置されることになるので、存在可能領域の境界線が45度に傾いた状態となる。また例えば上記実施例1のIDF平面図では、Y≧X−ln（N／N′）の領域にのみ各索引語が配置されることになるので、存在可能領域の境界線が45度に傾いた状態となる。

[0140] そこで、観察者が不慣れの場合でも、より的確に観察できるマップへの変換を与えるため、本実施例では、図11のマップ中の平行四辺形の端点A、B、C、及びDのそれぞれが長方形ABCDの四隅に来るように変換を施す。これにより、変換した横軸X′を専門性を表す軸、変換したY′を独創性を表す軸と解釈できれば、評価者が不慣れな場合でも、当該変換後のマップから、より的確に調査対象文書を評価できるようになる。

なお、図20のDF平面図のようにDF（P）の値を一律に文書数Nで除した場合でも、存在可能領域の境界線を45度より垂直に近づけることが可能であるが、却って原点付近への索引語座標の集中が顕著になるなど索引語座標の集中箇所が生じてしまう。そこで以下の変換例1ー3に示すように、横軸に沿った移動量が縦軸の値によって異なるような変換を施すことが望ましい。変換例1ー3におけるX値に対する変換は、Y値との関数によって与えられる。

[0141] ＜6−1．変換例1：図21（線形変換）＞

図21は、図11の条件のままで、図11の平行四辺形ABCDを長方形ABCDに変

換したものである。特に、Y＝X の直線に平行な線を、Y軸値を保ったままY軸と平行な線に変換したものである。すなわち、変換前の点の座標を(X, Y)とおくと、変換後の点の座標(X′ , Y′)は、

[数1]

$$(X' , Y') = (X - Y + \mathrm{const}, Y)$$

で表わされる。但し、式中でconst＝0のとき、図11の平行四辺形ABCDのうちの独創的着目語領域aはX′ ＜0 なる領域に変換されて収まる。一方、式中でconst＝β_2／2のとき、同領域はX′ ≧0 なる領域に変換されて収まる。図21はconst＝β_2／2の場合を示している。

[0142] 図21から、独創的着目語領域aに「欲求」「ホログラム」「絵」「プラスチック」「外面」などの特徴索引語を見つけ、専門語領域bには特徴索引語を見出せず、類似文書群規定語領域cに「コンテンツ」「編集」などの特徴索引語を見つけることができる。

[0143] 図21のように表わされたマップを調査対象文書の評価者が観察したとき、マップが図11などのように平行四辺形ではなく、長方形に分かれているので、特徴索引語をより的確に評価できる。

[0144] ＜6−2. 変換例2:図22(スケール変換)＞

図22は、図11の条件のままで、図11のX値を、Y軸から辺BCにかけてのX軸方向に沿った長さに対する比率で、スケール変換したものである。すなわち、変換前の点の座標を(X, Y)とおくと、変換後の点の座標(X′ , Y′)は,

[数2]

$$(X' , Y') = (X * (\alpha + \beta 2/2) / (Y + \alpha), Y)$$

で表わされる。これは一次双曲変換である

[数3]

$$(X' , Y') = (\mathrm{const} * X / (Y + \alpha), Y)$$

の特別な場合に相当する。

[0145] 図22から、独創的着目語領域aに「プラスチック」「外面」「ホログラム」「絵」などの特徴索引語を見つけ、同じく専門語領域bには、該当する特徴索引語を見つけることが

できず、又同じく類似文書群規定語領域cには、「コンテンツ」「編集」などの特徴索引語を見つけることができる。

[0146] 図22では、マップの左上方に索引語の不存在領域が残っているが、右方の存在領域の境界線は垂直になっている。従って、図22のように表わされたマップを調査対象文書の評価者が観察したとき、特に類似文書群規定語領域cの特徴索引語をより的確に評価できる。

[0147] ＜6−3. 変換例3:図23(下半部双曲変換)＞

図23は、図11の条件のままで、図の上半分の平行四辺形には変換例1の式を適用し、図の下半分には変換例2の式を適用して変換(複合変換)したものである。すなわち、変換前の点の座標を(X, Y)とおくと、変換後の点の座標(X′, Y′)は、

[数4]

$$X' = [X(\alpha+\beta 2/2)/(Y+\alpha)] \times \Theta(\beta 2/2-Y) + (X-Y+\beta 2/2) \times \Theta(Y-\beta 2/2)$$

但し、$x>0$のとき$\Theta(x)=1$、
$x=0$のとき$\Theta(x)=1/2$、
$x<0$のとき$\Theta(x)=0$

$$Y' = Y$$

で表わされる。

[0148] 図23から、独創的着目語領域aに「絵」「ホログラム」「外面」「プラスチック」「欲求」などの特徴索引語を見つけ、同じく専門語領域bには、該当する特徴索引語を見つけることができず、又同じく類似文書群規定語領域cには、「コンテンツ」「編集」などの特徴索引語を見つけることができる。

[0149] 図23では、マップ左右の索引語の不存在領域が解消され、境界領域がいずれもX軸に垂直になっている。従って、図23のように表わされたマップを調査対象文書の評価者が観察したとき、各領域の特徴索引語をより的確に評価できる。

[0150] 図24は、調査対象文書dとして「抗腫瘍剤」に関する公開特許公報を2件選び、図23と同じ方法で変換(複合変換)したときの、マップ表示の具体例である。

図24でも、図23と同様に、マップ左右の索引語の不存在領域が解消され、境界領域がいずれもX軸に垂直になっている。従って、各領域の特徴索引語をより的確に評価できる。

[0151] 図24には、独創的着目語領域a、専門語領域b、類似文書群規定語領域c、一般語領域dの存在位置を示す枠線が表示されている。このように各領域の存在位置をマップ上に表示することにより、各特徴索引語が属すべき領域をわかりやすく示すことができる。

各領域の存在位置の表示形態は枠線に限らず他の表示形態でも良いし、各領域の存在位置の表示に加えて「独創的着目語領域」等の具体的呼称を表示しても良い。また、枠線などにより各領域の存在位置をマップ上に表示することは、本実施例3のように座標値に対する変換を施す場合に限らず、他の実施例において行っても良い。

[0152] 各領域の存在位置をマップ上に表示して出力するには、例えば、各領域を示す枠線のみのデータを予め条件記録部310に保持しておき、マップ・リスト・コメント複合出力部440においてこれを読み出し、特徴索引語のマップ表示と重ね合わせて出力する。なお、処理すべきデータによってIDF(S)の上限値などが相違し、マップの大きさが異なる場合もあるので、得られるマップに合わせて枠線データの縦横長さを調整することが望ましい。また、本実施例3のように座標値に対する変換を施す場合は、そのような変換で得られる座標位置に適合した枠線データを予め準備しておくことが望ましい。

[0153] 図24から、独創的着目語領域aに「脆弱」「ユニーク」「集積」などの特徴索引語を見つけ、同じく専門語領域bには、「ZnPP」「ヘムオキシゲナーゼ」「プロトポルフィリン」などの特徴索引語を見つけ、又同じく類似文書群規定語領域cには、「腫瘍」「酵素」「細胞」などの特徴索引語を見つけることができる。

[0154] ＜6−4. 変換例4＞

上述の変換例以外にも、マップの観察を容易にする方法として、例えば、データを標準化する方法も可能である。すなわち、変換前の点の座標を(X, Y)とし、Xの平均をm(X)、Xの標準偏差をσ(X)としたとき(Yも同様とする)、変換後の点の座標(X′, Y′)を、

[数5]

$$(X' , Y') = ((X - m(X)) / \sigma(X), (Y - m(Y)) / \sigma(Y))$$

で与える。

この変換により、X及びYの平均値にX′ 軸及びY′ 軸が配置されるので、4領域への区分を容易にすることができる。

[0155] ＜7. 実施例4：自己組織化マップの応用＞

自己組織化マップ（SOM：Self-Organization Map）は、多数のデータを予備知識なしにクラスタリングできる技術である。このSOMの手法は、例えば、論文 Self-Organization Semantic Maps, H.Ritter and T.Kohonen, Biol. Cybern. 61(1989)241-254、或いは書籍 Self-Organizing Maps, T. Kohonen (Springer-Verlag, 1995) に開示されている。

[0156] 図25は、以下の説明の理解を容易にするために、図10を書き直した図である。図25において、各座標値は図11と同じ方法により得られた座標値である。同図において、点(0, $\beta_2/2$)をTとし、Tを通る傾き1の直線$Y=X+\beta_2/2$と直線BAの延長との交点をT′ とする。また、ADの中点をFとし、BCの中点をGとする。さらに、ABの中点をHとし、FGの中点をIとし、及びDCの中点をJとする。

[0157] 今、抽出された特徴索引語（キーワード）w_iがN_s個（$i=1, \cdots, N_s$）あるとする。これらN_s個の特徴索引語w_iは、平行四辺形ABCD内又は五角形BCDTT′ 内の領域に点在して分布する。しかし、これらの索引語がどの領域に属するのか、或いはどこにも属さないのか、一見して分類することは困難である。また、この平行四辺形は、斜めに傾いた形状であるので、評価者が特徴索引語の性格を即座に的確に読み取るのは困難である。

[0158] そこで、これら特徴索引語の座標点(X_i, Y_i)を、それらの性格をより簡単かつ的確に読み取ることが出来るような形態のマップ表示に変換した方がよい。その一つの手法として、この傾斜した平行四辺形の各頂点A、B、C、及びDに近い領域に分布する特徴索引語を、四つの領域に自動的に分けてマップ表現できれば、これら特徴索引語の性格が一目瞭然となり、従って、評価者が的確に特徴索引語の性格を読み取ることが出来る。このようなマップ表現を実現させる一つの手法として、SOMを応用し

た以下の変換方法を用いる。

[0159] ＜7−1．自己組織化マップの応用例1：図26、図27＞

上述したN_s個の特徴索引語の座標点(X_i, Y_i)をこのマッピング処理の入力ベクトル$K(w_i)$とする。このX−Y平面中に、参照点$U_j(w_i;t)$を、任意の個数だけ任意の座標値として、取る。但し、この応用例1では、j:0，1，2，3，4，5，6，7，8，9，10の11点を取って、参照点を11点斜方格子の座標点として考える。この11点の初期値を、それぞれ図25中のA，B，C，D，F，G，H，I，J，T，T′ に対応する座標値$(m1_j, m2_j)$とする。

[0160] 図26は、自己組織化マップの応用例1における参照点の初期値を一例として示す図である。応用例1のマップ作成条件において、図26に示すように、参照点$U_j(w_i;t)$の初期値は、j:0ー10に対応して、それぞれ、0(0，0)、1(α／2，0)、2(α，0)、3(α／2＋β_2／2，β_2／2)、4(α＋β_2／2，β_2／2)、5(α／2＋β_2，β_2)、6(β_1，β_2)、7(β_2／2，β_2／2)、8(β_2，β_2)、9(0，β_2／2)、10(β_2／2，β_2)とする。

[0161] 参照点の初期値を設定したら、入力ベクトル$K(w_i)$で与えられた各索引語w_iにつき、各入力点から最近接の参照点$U_j(w_i;t)$の座標を、下記の更新則により、各索引語w_iに近づくように値を更新する。なお、上記$U_j(w_i;t)$の括弧内は各索引語w_iに対する依存性及び更新ステップ数tへの依存性を示している。このような更新をT_F回例えば1000回繰り返す。

[0162] こうして各索引語w_iにつき更新された最終ステップの参照点$U_j(w_i;T_F)$に基づいて、写像$R_j=(r1_j(w_i), r2_j(w_i))$を与える。特に、最終ステップの参照点$U_j(w_i;T_F)$のうち、各索引語$w_i$の座標から最近接の参照点$U_j(w_i;T_F)$に基づいて与えられる写像$R_j$が、マップへの出力座標となる。

[0163] 更新則は、例えば次の通りとする。

[数6]

更新則	$U_j(w_i; t+1) = U_j(w_i; t) + h(t)(K(w_i) - U_j(w_i; t))$
学習係数	$h(t) = \kappa(t)\ \mathrm{Exp}[-\lvert R_c(w_i; t) - R_j(w_i; t)\rvert / (2\sigma(t)^2)]$
学習率	$\kappa(t) = 1 - t/T_F$
近傍サイズ	$\sigma(t) = \kappa(t)$
最近接参照点	$c = \mathrm{ArgMin}_j \lvert K(w_i) - U_j(w_i; t)\rvert$

[0164] 但し、tは更新ステップ回数の依存性を示す。また、$\delta_{\{j, 0\}}$は、クロネッカのδであり、$j=0$のとき$\delta_{\{j, 0\}}=1$、$j\neq 0$のとき$\delta_{\{j, 0\}}=0$を意味する。また、$\mathrm{ArgMin}_j(x)$は、xが最小となるjを返す関数である。なお、近傍サイズを、$\sigma(t)=\kappa(t)$としたのは、$\sigma(t)$の函数形の詳細な項がこの変換の出力結果に大きな影響を与えないので、簡略化が可能なためである。

[0165] このような条件において、U座標からR座標へと座標変換する。すなわち、$U_j(w_i; T_F) = (m1_j(w_i; T_F), m2_j(w_i; T_F))$を$R_j(w)_i = (r1_j(w)_i, r2_j(w)_i)$に変換する。この変換方法には何通りかの方法があるが、例えば、索引語の存在領域の境界線が垂直となるように、以下のようにして行う。すなわち、

[数7]

(1)全てのjに対して
$r2_j = m2_j(w_i; T_F)$
(2)j=0, 1, 2, 3, 4, 5, 6に対して
$r1_j = m1_j(w_i; T_F) - (1 - \delta_{\{j, 0\}}) * m2_j(w_i; T_F) + \gamma$
(3)j=7, 8に対して
$r1_j = m1_j(w_i; T_F) - m2_j(w_i; T_F) + \alpha/4 + \gamma$
(4)j=9, 10に対して
$r1_j = m1_j(w_i; T_F) - m2_j(w_i; T_F) + \beta_2/2 + \gamma$

但し、$\gamma = \beta_2 - \alpha$とおいた。

また、R_jのjは、$K(w_i)$と$U_j(w_i; T_F)$との距離が最も小さい値をとるjとする。また、上式で$r1_j < 0$となった場合は、$r1_j = 0$とするのが好ましい。

[0166] 上述の変換によって、最近接の参照点U_jに基づく写像R_jが、特徴索引語の座標値(X_i, Y_i)に基づき写像された新たな座標値(X', Y')となる。

[0167] このようなマップ形成条件としての、j個の参照点の座標値、更新ステップ数、更新則、学習係数及びU座標系からR座標系へ変換条件は、予め条件記録部に格納し

ておいて、入力装置からの指示により、条件記録部310から読み出して来て、上述のマップ作成の演算を行えば、最終的にIDF座標系の座標値は、R座標系の座標値へと写像される。このマップ作成の演算につき説明する。

[0168] 本実施例4の上述した変換処理は、特徴索引語抽出部180で行う。この変換処理を行うには、まず、入力装置2からの指令により、条件格納部310から更新則を読み出す。

[0169] 続いて、入力装置2からの指令により、実施例1と同様の抽出方法によって得られるIDF平面の座標系を作業結果格納部320から読み出してきて表示させる。表示画面を見ながら、IDF平面に分布している特徴索引語をN_s個指定して入力値を設定する。さらに、入力装置2からの指令により、更新回数T_Fを設定する。

[0170] これらの設定が終了すると、自動的に或いは入力装置からの演算開始指令により、マップ作成の演算を開始して、N_s個の特徴索引語の座標値(X_i, Y_i)は、最終的に、R座標の座標値へと写像される。

[0171] 図27は、一例として、図11の各座標点に対し上述の変換を行って得たマップを示す図である。図27からも理解出来るように、各座標点は、二つの直線a—a及びb—bによって分けられた4つの長方形の領域に分離されることがわかる。

[0172] ＜7−2. 自己組織化マップの応用例2:図28、図29＞

この変換は応用例1に類似する例である。応用例1では入力ベクトル$K(w_i)$として特徴索引語の座標点(X_i, Y_i)をそのまま用いたが、本応用例2では各座標点の値に対して予め変換を施し、入力ベクトルとして、

$$K(w_i) = (Y_i, Y_i - X_i + \alpha)$$

を用いる。

この変換により入力ベクトル$K(w_i)$はほぼ直線$Y = \alpha + \beta_2/2$、$X = \beta_2$、X軸及びY軸で囲まれる矩形領域内に分布することになる。そこで参照点の初期値もこの領域内に分布させる。

[0173] 図28は、応用例2で用いる参照点の配置例を示しており、これら11個の参照点に0番から10番までの番号を付して示してある。各参照点の初期値は、横軸上の$(\beta_2/6, 0)$ $(\beta_2/2, 0)$ $(5\beta_2/6, 0)$の各点をそれぞれ通る直線と及び縦軸上の$(0, \alpha$

／6)(0, α／2)(0, 5α／6)(0, α＋β_2／4)の各点を通る直線との11個の交点での座標値である。

[0174] そして、応用例1と同様の更新則に従い、各索引語w_iにつき参照点$U_j(w_i;t)$をT_F回更新する。

[0175] U座標からR座標($r1_j(w_i)$, $r2_j(w_i)$)への座標変換は、出力座標の存在点を直線X＝α＋β_2／2、Y＝β_2、Y軸及びX軸で囲まれる矩形領域内に分布させるように、すべてのjに対して、以下のようにして行う。

[数8]

$$r1_j(w_i) = \alpha + \beta 2/2 - m2_j(w_i;T_F) - \delta_{\{j,6\}}(\alpha/6 + \beta 2/4)$$
$$r2_j(w_i) = m1_j(w_i;T_F)$$

このような変換処理によって、最近接の参照点U_jに基づく写像R_jが、特徴索引語の座標値(X_i, Y_i)に基づき写像された新たな座標値(X′ , Y′)となる。

[0176] 図29は、一例として、図11の各索引語の座標点に対し、上述した変換処理を行った結果を示す図である。この変換処理によって得られた各座標点は、二つの直線a−a及びb−bによって分けられた四つの長方形の領域に分離されることが分かる。また、図27の新たな座標系の場合と同様に、図11の左上領域に示した空白領域に対応する空白領域が解消していることが分かる。

[0177] ＜7−3. 自己組織化マップの応用例3:図30、図31＞

この変換も応用例1に類似する例である。まず、図11の各索引語の座標値(X_i, Y_i)に対して、実施例3で説明したスケール変換を行って入力ベクトルK(w_i)とする。そしてこの例では、新たな16個の参照点を用いて、応用例1と同様の変換処理を行う。

[0178] 図30は、この16個の参照点を示しており、この座標系において、16個の参照点に0番から15番までの番号を付して示してある。各参照点の座標値は、横軸上に(β_1／8, 0)(3β_1／8, 0)(5β_1／8, 0)(7β_1／8, 0)の各点をそれぞれ通る直線と及び縦軸上の(0, β_2／8)(0, 3β_2／8)(0, 5β_2／8)(0, 7β_2／8)の各点を通る直線との16個の交点である。

[0179] この16点格子を用いた変換の場合には、入力ベクトルとして

[数9]

$K(w_i)=(X_i*(\alpha+\beta 2/2)/(Y_i+\alpha),\ Y_i)$

を用いることで、予めスケール変換を施し索引語の存在領域の境界線を垂直にしておく。そして、応用例1と同様の更新則に従い、各索引語w_iにつき$U_j(w_i;t)$をT_F回更新する。

[0180] U座標からR座標($r1_j(w_i)$, $r2_j(w_i)$)への座標変換は、すべてのjに対して、以下のように行う。

$r1_j(w_i)=m1_j(w_i;T_F)$

$r2_j(w_i)=m2_j(w_i;T_F)$

[0181] このような16点の参照値を用いた変換処理によって、最近接の参照点U_jに基づく写像R_jが、特徴索引語の座標値(X_i, Y_i)に基づき写像された新たな座標値(X', Y')となる。

[0182] 図31は、一例として、図11の各索引語の座標点に対し、上述した16点の参照値を用いた変換処理を行った結果を示す図である。この変換によって得た各座標点は、直線a2—a2と、直線b2—b2により分けられる4つの長方形の領域に配置されることが分かる。

[0183] ＜7—4. 自己組織化マップの応用例4:図32＞

この変換も応用例1に類似する例である。応用例1ー3では入力ベクトル$K(w_i)$及び参照点$U_j(w_i;t)$が2次元であったのに対し、本応用例では入力ベクトル及び参照点を、$2+N_s$次元とする。

[0184] まず、入力ベクトル$K(w_i)$は、特徴索引語の座標値(X_i, Y_i)と、当該特徴索引語とN_s個の特徴索引語の各々との共起度を用いたベクトルV_iを用いて、

$K(w_i)=(X_i, Y_i, V_i)$

で表現する。

ここで共起度ベクトルV_iは、共起度行列の成分Co(i, i′)から得られる共起データ$Co_{\{ii'\}}$(但し、i′ =1, 2, ・・・, N_s)を用いて、

$V_i=(Co_{\{i1\}}, Co_{\{i2\}}, \cdots, Co_{\{iNs\}})$

で表現されるN_s次元ベクトルとする。

[0185] ここで共起度行列の成分Co(i, i′)は、

[数10]

$$Co(i, i') = \sum_{\{sen \in d\}} TF(w_i, sen)^{\tau} \times TF(w_{i'}, sen)^{\tau} \times \mu_i \times \mu_{i'}$$

とする。TF(w, sen)はセンテンスsen中での索引語wの出現頻度、τは冪、μは重みを表す。ここでは例えばτ＝1／2、μ＝1を選ぶ。

TF(w, sen)は、センテンスsen内に索引語wが出現する場合は1以上の数となり、出現しない場合は0となるから、上記$TF(w_i, sen)^{\tau} \times TF(w_{i'}, sen)^{\tau} \times \mu_i \times \mu_{i'}$は、同一センテンスsen内に特徴索引語$w_i$と特徴索引語$w_{i'}$が共に出現する(共起する)場合は1以上の数となり、一方又は双方が出現しない(共起しない)場合は0となる。これを調査対象文書d内のすべてのセンテンスsenにつき合計したものが、共起度行列の成分Co(i, i′)である。

なお、τ＝1／2、μ＝1を選んだのは、共起度行列の対角成分Co(i, i)をTF(w_i, d)とするためである。

[0186] 共起度ベクトルV_iの成分である共起データ$Co_{\{ii'\}}$は、共起度行列の成分Co(i, i′)をi′に関する平均で標準化したものを、V_iの次元数N_sの平方根で除したものであり、以下のように表現される。

[数11]

$$Co_{\{i\,i'\}} = \frac{Co(i, i') - (1/Ns)\sum_{i'=1}^{Ns} Co(i, i')}{\sigma(Co(i, i')) \times \sqrt{Ns}}$$

ここで、$(1/N_s)\Sigma_{i'=1}^{Ns} Co(i, i')$は、Co(i, i′)のi′＝1, 2, ・・・, N_sに関する平均である。

また、σ(Co(i, i′))は、Co(i, i′)のi′＝1, 2, ・・・, N_sに関する標準偏差である。

このように共起度行列の成分Co(i, i′)を標準化し、かつ次元数N_sの平方根で除して共起度ベクトルV_iの成分$Co_{\{ii'\}}$を得ることにより、共起度ベクトルV_iの大きさは1となる。

[0187] 入力ベクトルとしては、上記$K(w_i) = (X_i, Y_i, V_i)$で表される$2+N_s$次元ベクトルのう

ち、X_iやY_iの部分については応用例2又は応用例3のような変換を施したものを用いてもよい。但し、ここでは上記$K(w_i)=(X_i, Y_i, V_i)$をそのまま用いるものとして説明する。

[0188] 次に参照点$U_j(w_i;t)$の初期値は、上記応用例1の参照点の初期値の座標$(m1_j, m2_j)$を用いて、

$(m1_j, m2_j, L_j)$

で表現する。ここでL_jはN_s次元ベクトルで、各成分は区間[0, 1]のランダム値をとるものとする。

[0189] 次に応用例1と同様に、入力ベクトル$K(w_i)$で与えられた各索引語w_iにつき、各入力点から最近接の参照点$U_j(w_i;t)$の座標をT_F回更新する。更新則も、例えば応用例1で用いた上記[数6]を用いる。

[0190] そして、各索引語w_iにつき更新された最終ステップの参照点$U_j(w_i;T_F)$のうち、各索引語w_iの入力ベクトルから最近接の参照点に基づいて、写像$R_j=(r1_j(w_i), r2_j(w_i))$を与える。U座標からR座標へと座標変換も、例えば応用例1で用いた上記[数7]を用いる。

ここで応用例1と異なる点は、応用例1では最終ステップの参照点$U_j(w_i;T_F)$が2次元であったのに対し、本応用例4では最終ステップの参照点$U_j(w_i;T_F)$は$2+N_s$次元である点である。しかし本応用例4においても、最終ステップの参照点$U_j(w_i;T_F)$のうち2つの成分$m1_j(w_i;T_F)$、$m2_j(w_i;T_F)$のみを用い、2次元の写像R_jを得るので、上記[数7]の変換式をそのまま用いることができる。こうして得られる写像R_jが、特徴索引語の座標値(X_i, Y_i)に基づき写像された新たな座標値(X', Y')となる。

[0191] 本応用例4では、入力ベクトルに共起度を用いた成分を加えているので、共起度の類似する特徴索引語w_i同士では参照点$U_j(w_i;t)$の更新過程が類似の挙動を示す。このため、R座標上に写像したときに、共起度の類似する特徴索引語同士は、共起度を考慮しない応用例1ー3のような場合に比べて近い位置に写像されることになる。

但し、本実施例の主目的は共起度又はその類似性そのものを示すことではなく、むしろIDF(P)とIDF(S)の関係を用いて調査対象文書の特徴を分析することに重きが

あるので、最終的な結果に共起度が及ぼす影響は小さくてよい。上記[数11]において共起度ベクトルV_iの各成分を求める際に次元数N_sの平方根で除したのはこのためである。なお、上記[数10]において$\tau = 1$としても良いが、このように次元数N_sの平方根で除しているので、$\tau = 1/2$の場合とあまり変わらない結果となる。

[0192] 図32は、一例として、図11の各索引語の座標点に対し、上述した共起度を加えた$2+N_s$次元ベクトルを用いた変換処理を行った結果を示す図である。この変換によって得た各座標点は、直線a−aと、直線b−bにより分けられる4つの長方形の領域に配置される。上記応用例1の結果である図27と比較すると、図27では例えば特徴索引語「料金」が一般語領域に分類され、特徴索引語「所望」が類似文書群規定語領域に分類されたのに対し、図32では特徴索引語「料金」が類似文書群規定語領域に分類され、特徴索引語「所望」が一般語領域に分類された。このように図32では、調査対象文書の特徴をより把握し易い分類が実現された。

[0193] ＜7−5．自己組織化マップの応用例5＞

上述の自己組織化マップの応用例1ー4により、各索引語がどの領域に属するかが明確になるので、そのデータを実施例1のような索引語リストやコメントの自動生成に用いることができる。例えば、自己組織化マップの応用例1ー4により得られた索引語のデータと、図12，図14，図16の索引語リストを生成するためのデータとをAND検索することにより、各領域に属する索引語を適切なものに絞り込むことができる。

[0194] なお、以上の実施例1ー4では、最も好ましい例として類似文書群Sを比較対象文書群Pの中から選出する場合を説明したが、類似文書群Sの選出元となる選出源文書群は、比較対象文書群P以外の文書群であってもよい。この場合、類似文書群Sは比較対象文書群Pの部分集合ではなくなるので、実施例3のスケール変換等をしても索引語の存在領域の境界線が垂直にはならない可能性がある。また、類似文書群Sを選出するための選出源文書群を、比較対象文書群Pとは別に入力する必要が生じる。しかし、それ以外は上記各実施例において説明したのと同様の作用及び効果を奏することができる。

[0195] ＜8. 実施例5:図33ー図37(索引語位置付けデータの集約)＞

次に、文書分布による文書特徴の分析及び文書群の性格付けについて説明する。実施例1ー4までは索引語分布による文書dの性格付けを行うものであるのに対し、本実施例は索引語情報(ミクロ情報)を文書情報(マクロ情報)に集約するとともに、調査対象を複数の文書からなる文書群に拡張する。調査対象文書群に含まれる調査対象文書の、他の文書群に対する大まかな位置付けや、調査対象文書群全体としての傾向を専門性や独創性といった観点から分析することのできる文書の特徴分析装置は、これまで知られておらず、本実施例はそれを実現するものである。

本実施例の文書特徴分析装置は、以下に説明する他は実施例1ー4の特徴索引語抽出装置と同様の構成を有する。以下では実施例1の特徴索引語抽出装置との相違点を主として説明する。

[0196] 特徴索引語のマップ上の分布により調査対象文書の性格を分析する代わりに、本実施例の文書特徴分析装置により大きな観測スケールを導入して、文書の分布により調査対象文書群を分析するには、次の置き換えを行なえばよい:

索引語 → 調査対象文書群の各文書;

索引語の(IDF(P), IDF(S))ベクトル → 調査対象文書群の各文書における索引語の(IDF(P), IDF(S))ベクトルの平均;

調査対象文書d → 調査対象文書群;

類似文書群S → 調査対象文書群と共通の属性を有する文書群である同類文書群S。

[0197] ここでは例として、調査対象文書群を1つの調査対象企業の文書群とし、同類文書群Sを当該企業と同じ業界に属する企業群の文書群とした場合について説明する。

本実施例でも特許文書を例にとると、例えば比較対象文書群Pを全特許文書群とし、調査対象企業と同業界に属する企業群の特許文書群である同類文書群Sを選出する。そして、調査対象企業の文書dについて、索引語のそれぞれにつきP及びSにおけるIDF演算をし、各文書dにおけるそれらの平均値などによる中心点を算出し、この値をもって各文書dのXY座標とする。当該企業の文書dの座標をXY平面にマップすると、当該企業の文書分布が得られる。

[0198] ＜8−1．実施例5の構成及び作用＞

図33は、実施例5の文書特徴分析装置のハードウェア構成を示す図である。図34は、当該装置の処理装置1の動作を示すフローチャートであり、図35は、当該装置の出力装置4におけるマップ出力の動作を示すフローチャートである。

[0199] 実施例1の類似文書群Sと異なり実施例5の同類文書群Sは類似度に基づいて選出されるものではない。よって図33に示すように、図2の類似度演算部150は不要であり、従って図2のTF(d)演算部121、TF(P)演算部141も不要である。同様に、図34に示すように、図4の類似度演算ステップS150は不要であり、図4のTF(d)演算ステップS121、TF(P)演算ステップS141も不要である。

[0200] 同類文書群Sの選出は、入力装置2の抽出条件その他入力部230で入力された条件に従い、例えば以下のように行う。すなわち、業界分類から調査対象企業と同業界の企業を検索する場合は、まず条件記録部310に、主要企業名及びそれらの「標準産業分類」又はその他の業界分類を記憶させておく。そして、同業界企業検索部155により調査対象企業と同業界に属する企業名を検索する。検索された企業名をキーとして、同類文書群S選出部160が比較対象文書群Pの書誌データを対象に検索することで、同類文書群Sを選出する。

なお、同類文書群S選出部160は、上記同業界の文書群から更に一定条件で絞り込んで、同類文書群Sとしても良い。

[0201] 同類文書群S選出部160は、こうして選出された同類文書群Sを索引語(S)抽出部170等に出力する。索引語(S)抽出部170は、同類文書群Sの入力を受けたら索引語(S)を抽出し、IDF(S)演算部171等に送る。IDF(P)演算部142及びIDF(S)演算部171の演算結果をもとに、中心点算出部173で中心点の算出を行う。

[0202] また、実施例5は文書分布マップの出力を主目的としている。実施例1のようなリスト出力を行わない場合は、図33に示すように図2のリスト出力条件読み出し部420、リスト用データ取り込み部422は不要である。同様に、図35に示すように図5のリスト出力条件読み出しステップS420からリスト生成ステップS423までの各ステップも不要である。実施例1のようなコメント出力を行わない場合は、図2のコメント追記条件読み出し部430、コメント追記部432は不要である。同様に、図5のコメント追記条件読み

出しステップS430からコメント生成ステップS433までの各ステップも不要である。

[0203] 調査対象企業の各文書における中心点の座標値は、各索引語w_iの座標値に、TF重率:

$$\rho(w_i) = TF(w_i;d) / \Sigma\ TF(w_i;d)$$

で重み付けをした平均値であることが望ましいが、これに限られるものではなく単純平均値を用いてもよい。

[0204] 調査対象企業の文書数が膨大にある場合、代表的な文書に絞ってマップに出力する方が調査対象企業の文書群としての傾向を把握し易くする上で好ましい。そこで、調査対象文書群の中から、当該調査対象文書群に対して類似性の高い文書と、当該調査対象文書群に対して類似性の低い文書とを文書抽出部180にて抽出して出力する。

[0205] 調査対象文書群に対する各文書の類似性の判定は、例えば、各文書dにつき、各索引語w_iで調査対象文書群を検索したときのヒット文書数$DF(w_i, E0)$の平均値$(1/d_N)\{DF(w_1, E0)+DF(w_2, E0)+\cdots+DF(w_{dN}, E0)\}$を算出し、この平均値の高い文書($d_N$は当該文書d内の索引語数)を「類似」、低い文書を「非類似」とする。抽出の方法としては、例えば上記平均値の昇順及び降順の一定数を抽出する方法、また例えば上記平均値を調査対象文書群の文書数で除したものをZとしたときに、「全Zの平均値＋全Zの標準偏差」以上のZをとる文書と、「全Zの平均値−全Zの標準偏差」以下のZをとる文書とを抽出する方法などが考えられる。

[0206] ここで述べた類似性の判定による代表的文書への絞り込みは、調査対象文書群を絞り込むのに使うほか、同類文書群Sを選出する際の絞り込みにも使うことができる。すなわち、同業界の文書群の各文書につき、各索引語で上記同業界の文書群を検索したときのヒット文書数の平均値を算出し、この平均値の高い(類似)文書及び低い(非類似)文書に絞り込んで同類文書群Sとして選出する。なお、同類文書群Sを選出する際の絞り込みは、類似性の判定によるほか、同業界の文書群から無作為抽出することにより行っても良いし、IPCで絞り込んでも良い。

[0207] ＜8−2. マップ出力例＞

図36は、調査対象文書群である企業1社の全文書のうち、類似性の高い文書20

件、及び類似性の低い文書20件について、業界内における位置付けによる文書特徴を示した図である。この図は、本発明における企業の文書特徴表現図に相当する。図36では、各文書の中心値として、単純平均値を使った。該企業の文書dをIDF平面図にマップすると、企業の文書の分布が得られる。

[0208] こうして得られたマップでは、直線Y＝(β_2／β_1)Xより上の領域に、殆どの文書の座標が分布する(β_1は比較対象文書群Pの文書数Nに基づくX座標の最大値lnN、β_2は同類文書群Sの文書数N′に基づくY座標の最大値lnN′である)。そのうちY＝Xより左上の領域には独創的着目語の多い文書が現れ、X＝$\beta_1-\beta_2$より右の領域には専門語の多い文書が現れる。その間の領域には標準的な文書が現れるので、どの領域に文書が多く分布するかにより、企業の文書の傾向を把握することができる。

[0209] Y＝Xより左上の領域に現れる文書が独創的着目語の多い文書であると評価できる理由を説明する。同類文書群Sに大量の文書を加える時のDF値変化は、DF値の増加率が文書数の増加率と同等であるものと、DF値が殆ど変化しないものと、DF値が急激に増加するものと、の三種類に分類される。それぞれの場合のIDF変化は、順に、変化なし、増加、減少、となるので、同類文書群Sに大量の文書を加えたときのIDF平面上の索引語分布は、直線Y＝X方向へ移動し易い傾向を持つ。ここでは各文書での平均をとっているので、一層直線Y＝X方向へと近づく傾向が現れる。この傾向は、Y＝Xより上方の領域には独創的着目語の多い文書が現れるということを示唆する。

また、X＝$\beta_1-\beta_2$より右の領域に現れる文書が専門語の多い文書であると評価できる理由を説明する。類似文書群規定語領域cの索引語座標と一般語領域dに属する索引語座標の平均をとった場合、類似文書規定語領域cの端点C($\beta_1-\beta_2$, 0)のX座標値がおよそ最大値であると考えられる。従って、X＝$\beta_1-\beta_2$より右の領域には標準的な文書は現れず、専門語の多い文書であると評価できる。

以上より、残りのY≦Xで且つX≦$\beta_1-\beta_2$の領域が、標準的な文書の領域となる。

[0210] また、直線Y＝(β_2／β_1)Xより上の領域に、殆どの文書の座標が分布する理由を説明する。各文書の中心値の座標は索引語の平均値をとっていることから、一様性

の仮定(DF(P)＝N／k、DF(S)＝N′／k、k≧1)をとることができる。この一様性の仮定と平面座標の定義(X, Y)＝(<IDF(P)>$_w$, <IDF(S)>$_w$)とから、Y＝(β_2／β_1)X＋(α／β_1)lnkが導かれる。これより、k≧1を満たすkに対してY≧(β_2／β_1)Xが成立する。

[0211] 以上説明した傾向によれば、本実施例の文書特徴分析装置を利用して、人間が調査対象文書群や同類文書群などの内容を一切読むことなく、調査対象文書の大まかな位置付けや傾向を分析することができる。すなわち、調査対象文書群である企業の文書群のうち、特定の文書が業界において標準的な文書か、専門的性格を持つ文書か、或いは独創的性格を持つ文書かを知ることができる。また、調査対象文書群である企業の文書群のうち、標準的な文書を検出したり、専門的性格を持つ文書を検出したり、又は独創的性格を持つ文書を検出したりすることもできる。更に、調査対象文書群全体としての傾向を、標準的文書の多い文書群、独創的性質を持つ文書の多い文書群、或いは専門的性質を持つ文書の多い文書群というように評価することができる。

[0212] また、図36では、調査対象文書群のうち、類似性の高い文書20件、及び類似性の低い文書20件を抽出してマップ出力している。このような抽出により、調査対象文書群に対する類似性が低く、且つ同類文書群Sに対する独創的性質又は専門的性質の高い文書は、特に独自性の高い文書だという評価をすることも可能である。また、調査対象文書群に対する類似性が低くても、同類文書群Sに対しては独創的性質又は専門的性質の低い、或いは標準的な文書は、既成概念や公知技術の組合せの可能性があるという評価も可能である。

[0213] 図37は、調査対象文書群として同業界に属する3社の文書群を選び、各社の文書特徴を示した図である。これらを比較すると、A社、C社の文書には専門語の多い文書が多い傾向が見られ、B社の文書には独創的着目語の多い文書が多い傾向が見られる。この図は、本発明における企業の文書特徴表現図に相当する。このように調査対象文書群として複数の文書群を分析し、文書群相互の比較をすることで、文書群全体としての傾向をより的確に評価することができる。

[0214] ＜8−3. 実施例5の変形例1(同類文書群の選出)＞

以上の例では同類文書群Sとして調査対象企業と同業界に属する企業の文書群又はこれを更に絞り込んだ文書群を用いた場合を説明したが、同類文書群Sはこれに限られるものではない。例えば、調査対象企業の文書群と同分野に属する文書群をIPCなどにより検索して同類文書群Sとしても良い。

[0215] IPCにより同分野に属する文書群を検索する場合は、図33の処理装置1において、図示しないIPC抽出部を設け、このIPC抽出部により、調査対象企業の全特許文書の書誌データからIPCを抽出する。多数のIPCが抽出される場合は、該当文書数の最も多いIPCを、上位所定数のみ抽出する。そして、抽出されたIPCをキーとして、同類文書群S選出部160が比較対象文書群Pの書誌データを対象に検索することで、同類文書群Sを選出する。かかる選出条件は、例えば入力装置2の抽出条件その他入力部230で入力する。

[0216] こうして選出された同類文書群Sを用いることにより、調査対象企業の文書群について同一分野内の文書群における位置付けや傾向を分析することができる。

[0217] ＜8−4. 実施例5の変形例2（調査対象文書群の取り方1）＞

以上の例では調査対象文書群として調査対象企業の文書群を用いた場合を説明したが、調査対象文書群はこれに限られるものではない。例えば、不特定多数の特許文書群のうち同分野に属する文書群をIPCなどにより検索して調査対象文書群としても良い。

[0218] 例えば、調査対象文書群として、2000年に特許出願され、あるIPCを付与された文書群を分析する場合を考える。同類文書群Sとして、例えば1980ー1999年に特許出願され、上記IPCと同じIPCを付与された文書群を選出する。他の条件は上記と同じとして調査対象文書群を分析する。

[0219] これにより、当該IPCを付与された技術分野における2000年の出願動向が、過去20年と比べて独創的な方向にシフトしたのか、専門的な方向にシフトしたのか、標準的と言える範囲にとどまるのか、を評価することができる。また、当該IPCを付与された技術分野における2000年の出願のうち、特定の出願が過去20年の出願に対して独創的性質を持つのか、専門的性質を持つのか、標準的と言える範囲にとどまるのか、を評価することができる。また、当該IPCを付与された技術分野における2000年の出

願の中から、過去20年の出願に対して独創的性質を持つ出願、専門的性質を持つ出願、或いは標準的と言える出願を検出することもできる。

[0220] 更に、当該IPCを付与された技術分野における2000年の出願の分析結果を、他の調査対象文書群を用いた分析結果と比較することもできる。

例えば、調査対象文書群及び同類文書群Sの出願時期を上記と同じ2000年及び1980ー1999年とし、別のIPCについて同様の分析を行う。こうして異なるIPC同士で比較することにより、技術の入れ替わりが激しい分野、成熟した分野等の評価をすることができる。

また例えば、調査対象文書群として、2001年に特許出願され、あるIPCを付与された文書群を用い、同類文書群Sとして、1981年ー2000年に特許出願され、上記IPCと同じIPCを付与された文書群を用いて分析する。この分析結果と、上記2000年を調査対象とした場合の分析結果とを比較する。これにより、同一技術分野における2000年の出願動向と2001年の出願動向とを比較することもできる。

[0221] ＜8−5. 実施例5の変形例3（調査対象文書群の取り方2）＞

また例えば、調査対象文書群として、あるIPC（例えばサブグループまで指定：A61K6/05など）を付与された文書群を分析する場合を考える。同類文書群Sとして、当該IPCの上位階層に相当するIPC（例えばメイングループまで指定：A61K6/など）を付与された文書群を選出する。他の条件は上記と同じとして調査対象文書群を分析する。

[0222] これにより、調査対象文書群のうち特定の文書が、上位階層IPCの文書群に対して特異な性質（独創語が多い、専門語が多い等）を有する文書なのか、或いは標準的と言える範囲にとどまる文書なのかを評価することができる。また、調査対象文書群の中から、上位階層IPCの文書群に対して特異な性質（独創語が多い、専門語が多い等）を有する文書を検出し、或いは標準的な性質を有する文書を検出することができる。

請求の範囲

[1] 調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群の選出元となる選出源文書群、を入力する入力手段と、

前記調査対象文書内の索引語を抽出する索引語抽出手段と、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出手段と、

前記調査対象文書のデータに基づき、前記選出源文書群の中から前記類似文書群を選出する類似文書群選出手段と、

前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出手段と、

各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記類似文書群における出現頻度の関数値との組合せに基づき、各索引語とその位置づけデータとを出力する出力手段と、

を備えた、索引語抽出装置。

[2] 請求項1において、

前記選出源文書群として前記比較対象文書群を用いる、索引語抽出装置。

[3] 請求項1又は請求項2において、

前記類似文書群選出手段は、

前記調査対象文書及び前記選出源文書群の各文書について、当該文書に含まれる各索引語の当該文書における出現頻度の関数値又は各索引語の前記選出源文書群における出現頻度の関数値を成分とするベクトルを算出し、

前記調査対象文書について算出された前記ベクトルに対する類似度合いの高いベクトルをもつ文書を前記選出源文書群から選出して、類似文書群とする、索引語抽出装置。

[4] 請求項1乃至請求項3の何れか一項において、

前記出力手段は、各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の低い第1グループの索引語と、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの

索引語と、前記第1グループの索引語よりも前記類似文書群における出現頻度が高い第3グループの索引語と、をそれぞれ出力する、索引語抽出装置。

[5] 請求項1乃至請求項3の何れか一項において、

前記出力手段は、各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の高い第4グループの索引語よりも、前記比較対象文書群における出現頻度が低い第3グループの索引語と、前記第4グループの索引語よりも前記類似文書群における出現頻度が低い第2グループの索引語と、前記第3グループの索引語よりも前記類似文書群における出現頻度が低く且つ前記第2グループの索引語よりも前記比較対象文書群における出現頻度が低い第1グループの索引語と、をそれぞれ出力する、索引語抽出装置。

[6] 調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群、を入力する入力手段と、

前記調査対象文書内の索引語を抽出する索引語抽出手段と、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出手段と、

前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出手段と、

各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の低い第1グループの索引語と、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの索引語と、前記第1グループの索引語よりも前記類似文書群における出現頻度が高い第3グループの索引語と、をそれぞれ出力する出力手段と、

を備えた、索引語抽出装置。

[7] 調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群、を入力する入力手段と、

前記調査対象文書内の索引語を抽出する索引語抽出手段と、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出手段と、

前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出手段と、

各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の高い第4グループの索引語よりも、前記比較対象文書群における出現頻度が低い第3グループの索引語と、前記第4グループの索引語よりも前記類似文書群における出現頻度が低い第2グループの索引語と、前記第3グループの索引語よりも前記類似文書群における出現頻度が低く且つ前記第2グループの索引語よりも前記比較対象文書群における出現頻度が低い第1グループの索引語と、をそれぞれ出力する出力手段と、

を備えた、索引語抽出装置。

[8] 請求項1乃至請求項7の何れか一項において、

前記比較対象文書群又は前記類似文書群における出現頻度の関数値は、当該出現頻度の逆数に、前記比較対象文書群又は前記類似文書群の総文書数を乗じたものの対数である、索引語抽出装置。

[9] 請求項1乃至請求項8の何れか一項において、

前記出力手段は、

前記比較対象文書群における出現頻度の関数値を座標の第1軸にとり、

前記類似文書群における出現頻度の関数値を前記座標の第2軸にとって、前記索引語を配置し出力する、索引語抽出装置。

[10] 請求項4乃至請求項8の何れか一項において、

前記出力手段は、前記第1グループの索引語と、前記第2グループの索引語と、前記第3グループの索引語とを、それぞれリストして出力する、索引語抽出装置。

[11] 請求項4乃至請求項8の何れか一項において、

前記出力手段は、前記第1グループの索引語と、前記第2グループの索引語と、前記第3グループの索引語とを用いて、当該調査対象文書の解説文を自動生成して出力する、索引語抽出装置。

[12] 請求項1乃至請求項8の何れか一項において、

前記類似文書群の各文書は、前記比較対象文書群に含まれており、

前記出力手段は、前記比較対象文書群における出現頻度の関数値を、さらに変換して座標の第1軸にとり、前記類似文書群における出現頻度の関数値を前記座標の第2軸にとって、前記索引語を配置し出力するものであって、

前記変換は、前記類似文書群が前記比較対象文書群の部分集合であることによる、前記索引語の前記座標上における存在可能領域の境界線が、前記第1軸と垂直に近づくように変換するものである、索引語抽出装置。

[13] 請求項12において、

前記変換は、前記類似文書群における出現頻度との関数によって与えられる変換である、索引語抽出装置。

[14] 請求項1乃至請求項13の何れか一項において、

前記調査対象文書内の各索引語の、当該調査対象文書における出現頻度を算出する索引語頻度算出手段を更に備え、

前記出力手段は、前記調査対象文書内の各索引語の当該調査対象文書における出現頻度を反映して出力する、索引語抽出装置。

[15] 請求項1乃至請求項8の何れか一項において、

前記出力手段は、各索引語につき、

前記比較対象文書群における出現頻度の関数値を座標の第1軸にとり、

前記類似文書群における出現頻度の関数値を前記座標の第2軸にとった場合に、

前記座標上の複数の基準点のうち当該索引語に最も近い基準点に更に近づくように配置して座標上に出力する、索引語抽出装置。

[16] 請求項1乃至請求項8の何れか一項において、

座標上に複数の基準点の座標を設定する基準点設定手段と、

各索引語につき、前記比較対象文書群における出現頻度の関数値を座標の第1軸にとり、前記類似文書群における出現頻度の関数値を前記座標の第2軸にとった場合に、前記複数の基準点のうち当該索引語に最も近い基準点の座標データを、当該索引語に更に近づくように、所定回数にわたり更新する手段と、

前記更新された基準点に基づいて、当該索引語を配置する座標を算出する座標算出手段と、

を更に備え、

前記出力手段は、前記座標算出手段により算出された座標に基づいて、各索引語を前記座標に配置して出力する、索引語抽出装置。

[17] 調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群の選出元となる選出源文書群、を入力する入力ステップと、

前記調査対象文書内の索引語を抽出する索引語抽出ステップと、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出ステップと、

前記調査対象文書のデータに基づき、前記選出源文書群の中から前記類似文書群を選出する類似文書群選出ステップと、

前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出ステップと、

各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記類似文書群における出現頻度の関数値との組合せに基づき、各索引語とその位置づけデータとを出力する出力ステップと、

を備えた、索引語抽出方法。

[18] 調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群、を入力する入力ステップと、

前記調査対象文書内の索引語を抽出する索引語抽出ステップと、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出ステップと、

前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出ステップと、

各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の低い第1グループの索引語と、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの索引語と、前記第1グループの索引語よりも前記類似文書群における出現頻度が高い第3グループの索

引語と、をそれぞれ出力する出力ステップと、

を備えた、索引語抽出方法。

[19] 調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群の選出元となる選出源文書群、を入力する入力ステップと、

前記調査対象文書内の索引語を抽出する索引語抽出ステップと、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出ステップと、

前記調査対象文書のデータに基づき、前記選出源文書群の中から前記類似文書群を選出する類似文書群選出ステップと、

前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出ステップと、

各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記類似文書群における出現頻度の関数値との組合せに基づき、各索引語とその位置づけデータとを出力する出力ステップと、

をコンピュータに実行させる、索引語抽出プログラム。

[20] 調査対象文書、前記調査対象文書と比較される比較対象文書群、前記調査対象文書に類似する類似文書群、を入力する入力ステップと、

前記調査対象文書内の索引語を抽出する索引語抽出ステップと、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値を算出する第1出現頻度算出ステップと、

前記抽出された索引語の、前記類似文書群における出現頻度の関数値を算出する第2出現頻度算出ステップと、

各算出手段の結果に基づき、前記比較対象文書群においても前記類似文書群においても出現頻度の低い第1グループの索引語と、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの索引語と、前記第1グループの索引語よりも前記類似文書群における出現頻度が高い第3グループの索引語と、をそれぞれ出力する出力ステップと、

をコンピュータに実行させる、索引語抽出プログラム。

[21] 調査対象文書内の索引語について、

前記調査対象文書と比較される比較対象文書群における出現頻度の関数値を座標の第1軸にとり、

前記調査対象文書に類似する類似文書群における出現頻度の関数値を前記座標の第2軸にとって配置した、調査対象文書の性格表現図。

[22] 調査対象文書内の索引語を配置した、調査対象文書の性格表現図であって、

第1エリアに、前記調査対象文書と比較される比較対象文書群においても、前記調査対象文書群に類似する類似文書群においても、出現頻度の低い第1グループの索引語を配置し、

第2エリアに、前記第1グループの索引語よりも前記比較対象文書群における出現頻度が高い第2グループの索引語を配置し、

第3エリアに、前記第1グループの索引語よりも前記類似文書群における出現頻度が高い第3グループの索引語を配置した、調査対象文書の性格表現図。

[23] 調査対象文書内の索引語を配置した、調査対象文書の性格表現図であって、

第3エリアに、前記調査対象文書と比較される比較対象文書群においても前記調査対象文書群に類似する類似文書群においても出現頻度の高い第4グループの索引語よりも、前記比較対象文書群における出現頻度が低い第3グループの索引語を配置し、

第2エリアに、前記第4グループの索引語よりも前記類似文書群における出現頻度が低い第2グループの索引語を配置し、

第1エリアに、前記第3グループの索引語よりも前記類似文書群における出現頻度が低く且つ前記第2グループの索引語よりも前記比較対象文書群における出現頻度が低い第1グループの索引語を配置した、調査対象文書の性格表現図。

[24] 複数の調査対象文書を含む調査対象文書群、各調査対象文書と比較される比較対象文書群、前記調査対象文書群と共通の属性を有する同類文書群、を入力する

入力手段と、

前記各調査対象文書内の索引語を抽出する索引語抽出手段と、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値、を算出する第3出現頻度算出手段と、

前記抽出された索引語の、前記同類文書群における出現頻度の関数値、を算出する第4出現頻度算出手段と、

各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記同類文書群における出現頻度の関数値との組合せに基づき、前記各調査対象文書における中心点を算出する中心点算出手段と、

前記各調査対象文書における前記中心点のデータを出力する出力手段と、

を備えた、文書特徴分析装置。

[25] 請求項24において、

各調査対象文書における前記中心点の算出は、

各索引語についての、前記比較対象文書群における出現頻度の関数値と前記同類文書群における出現頻度の関数値とに基づく各索引語の座標値に、当該文書内の索引語頻度値合計に対する各索引語の索引語頻度値の比で重み付けをした平均値である索引語座標の加重平均値を算出することによって行う、文書特徴分析装置。

[26] 請求項24又は請求項25において、

前記調査対象文書群のうち、当該文書群に対して類似性の高い文書と、当該文書群に対して類似性の低い文書とを抽出して前記中心点のデータを出力する、文書特徴分析装置。

[27] 複数の調査対象文書を含む調査対象文書群、各調査対象文書と比較される比較対象文書群、前記調査対象文書群と共通の属性を有する同類文書群、を入力する入力ステップと、

前記各調査対象文書内の索引語を抽出する索引語抽出ステップと、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値、を算出する第3出現頻度算出ステップと、

前記抽出された索引語の、前記同類文書群における出現頻度の関数値、を算出する第4出現頻度算出ステップと、

各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記同類文書群における出現頻度の関数値との組合せに基づき、前記各調査対象文書における中心点を算出する中心点算出ステップと、

前記各調査対象文書における前記中心点のデータを出力する出力ステップと、

を備えた、文書特徴分析方法。

[28] 複数の調査対象文書を含む調査対象文書群、各調査対象文書と比較される比較対象文書群、前記調査対象文書群と共通の属性を有する同類文書群、を入力する入力ステップと、

前記各調査対象文書内の索引語を抽出する索引語抽出ステップと、

前記抽出された索引語の、前記比較対象文書群における出現頻度の関数値、を算出する第3出現頻度算出ステップと、

前記抽出された索引語の、前記同類文書群における出現頻度の関数値、を算出する第4出現頻度算出ステップと、

各索引語についての、前記算出された前記比較対象文書群における出現頻度の関数値と前記同類文書群における出現頻度の関数値との組合せに基づき、前記各調査対象文書における中心点を算出する中心点算出ステップと、

前記各調査対象文書における前記中心点のデータを出力する出力ステップと、

をコンピュータに実行させる、文書特徴分析プログラム。

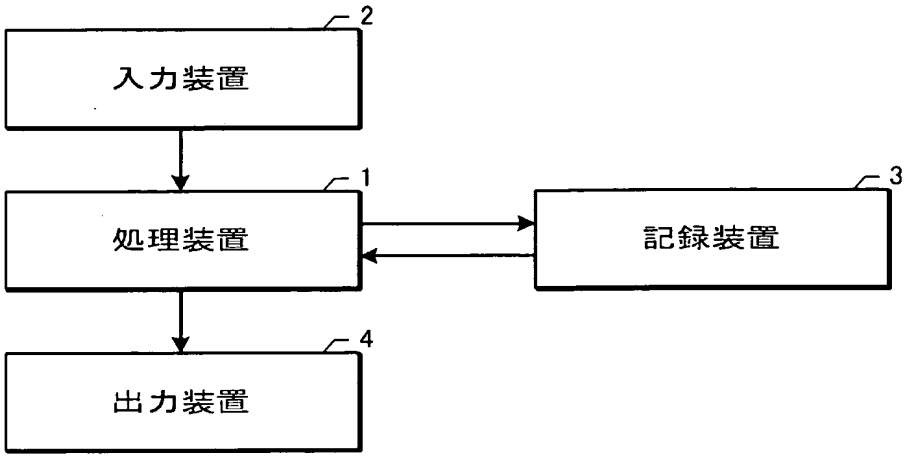
[29] 調査対象文書群に含まれる複数の調査対象文書について、各調査対象文書と比較される比較対象文書群に対する位置づけを座標の第1軸にとり、前記調査対象文書群と共通の属性を有する同類文書群に対する位置づけを前記座標の第2軸にとって配置した、調査対象文書の文書特徴表現図であって、

前記座標における前記各調査対象文書の座標値は、

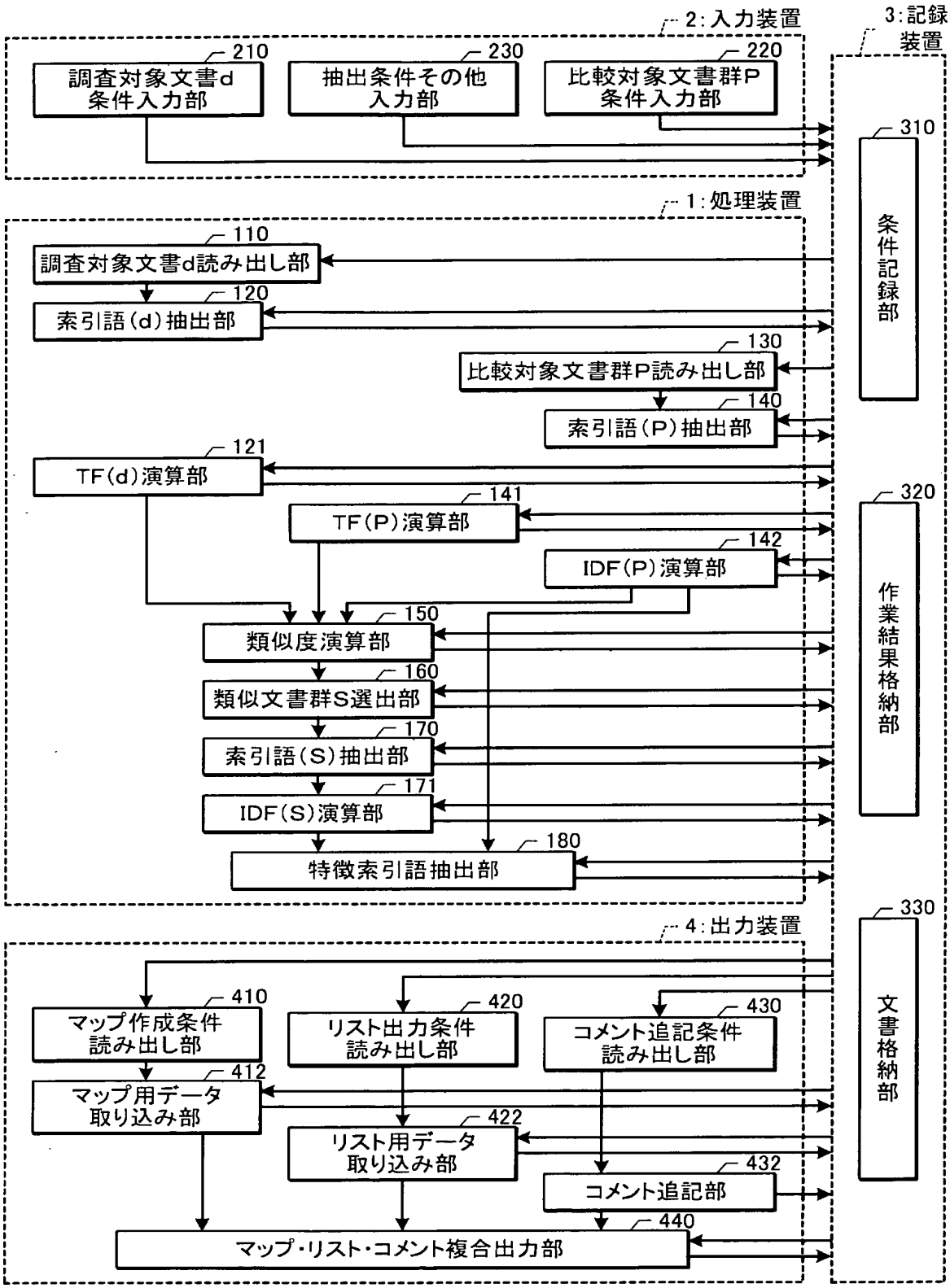
各調査対象文書内の各索引語の前記比較対象文書群における出現頻度の関数値と、各索引語の前記同類文書群における出現頻度の関数値と、を成分とする索引語座標値の、各調査対象文書における中心点とした、調査対象文書の文書特徴表

現図。

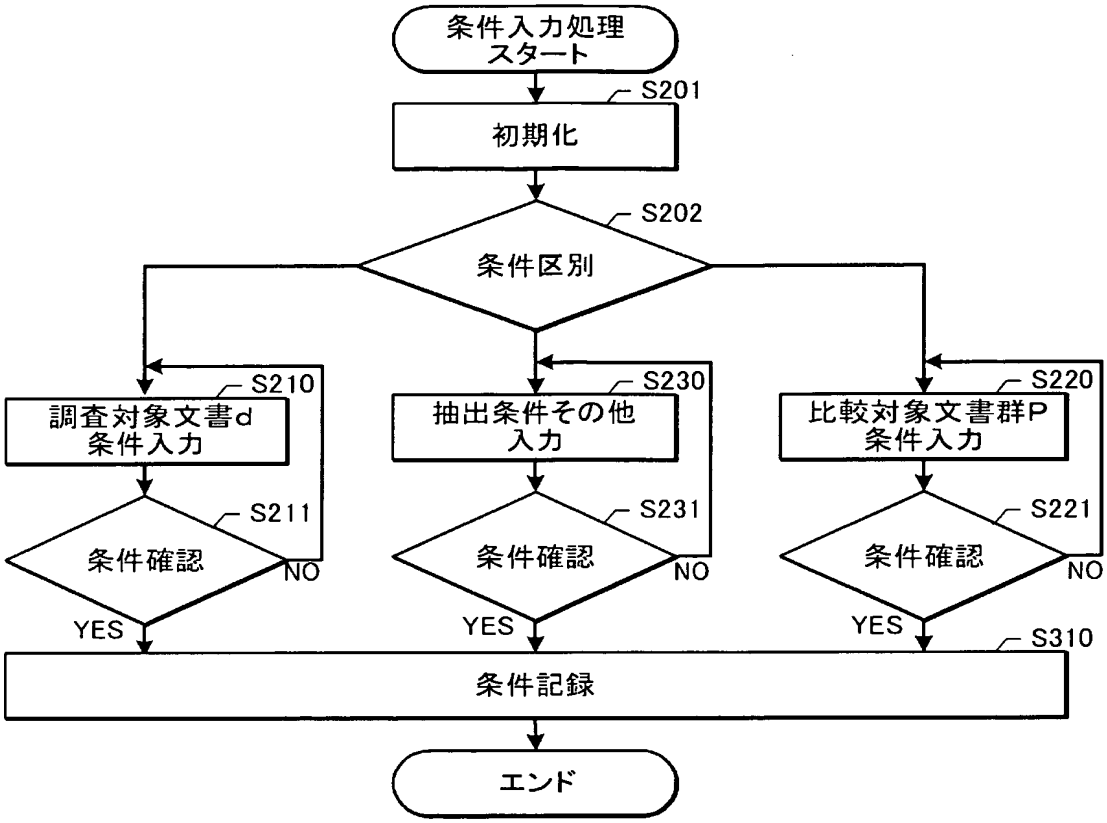
[図1]



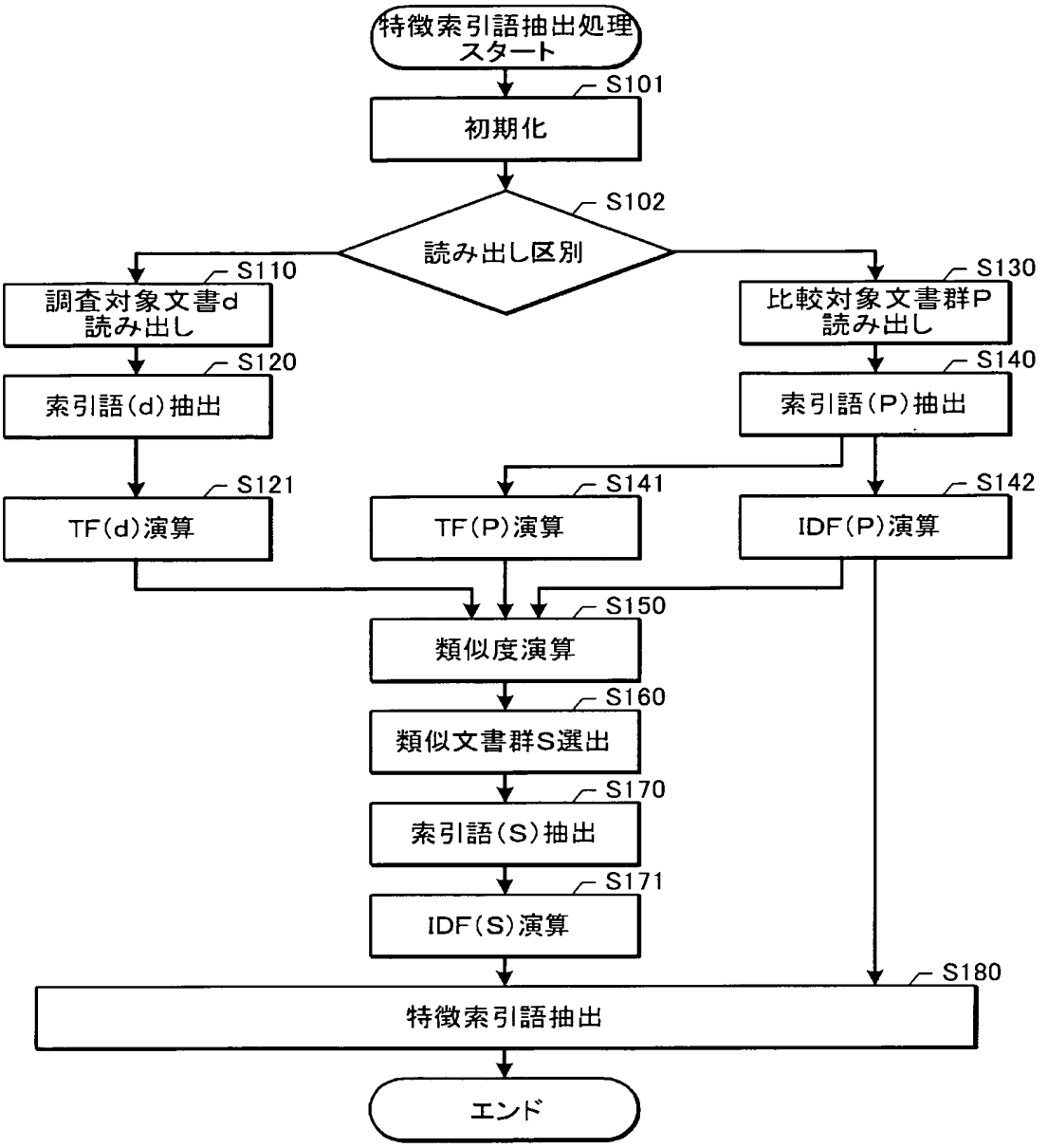
[図2]


2:入力装置
3:記録装置
210
調査対象文書d条件入力部
230
抽出条件その他入力部
220
比較対象文書群P条件入力部
310
条件記録部
1:処理装置
110
調査対象文書d読み出し部
120
索引語(d)抽出部
130
比較対象文書群P読み出し部
140
索引語(P)抽出部
121
TF(d)演算部
141
TF(P)演算部
142
IDF(P)演算部
320
作業結果格納部
150
類似度演算部
160
類似文書群S選出部
170
索引語(S)抽出部
171
IDF(S)演算部
180
特徴索引語抽出部
4:出力装置
330
文書格納部
410
マップ作成条件読み出し部
420
リスト出力条件読み出し部
430
コメント追記条件読み出し部
412
マップ用データ取り込み部
422
リスト用データ取り込み部
432
コメント追記部
440
マップ・リスト・コメント複合出力部

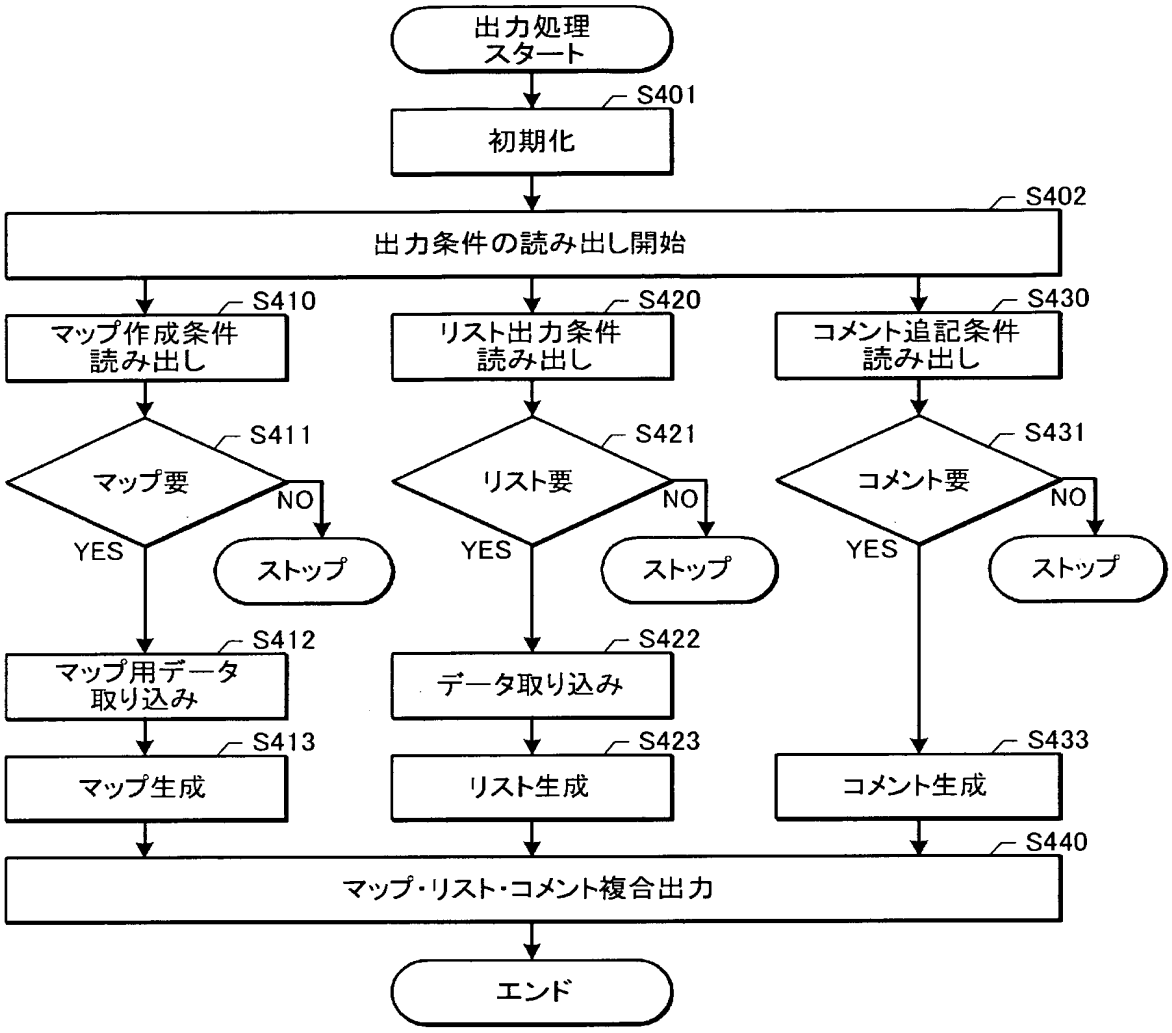
[図3]


条件入力処理
スタート
S201
初期化
S202
条件区別
S210
調査対象文書d
条件入力
S211
条件確認
NO
YES
S230
抽出条件その他
入力
S231
条件確認
NO
YES
S220
比較対象文書群P
条件入力
S221
条件確認
NO
YES
S310
条件記録
エンド

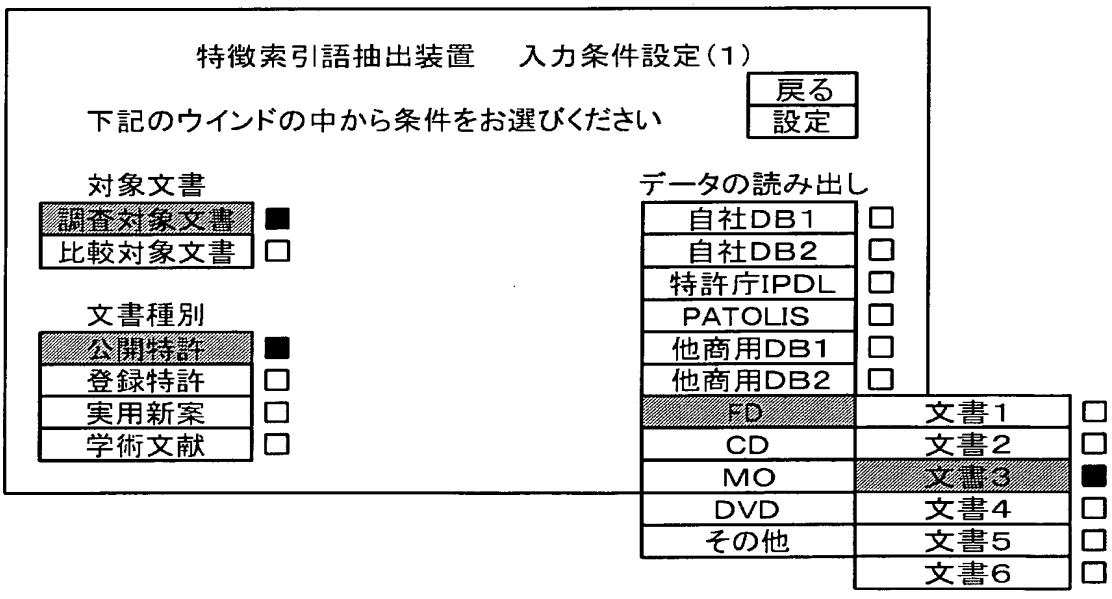
[図4]


特徴索引語抽出処理
スタート
S101
初期化
S102
読み出し区別
S110
調査対象文書d
読み出し
S120
索引語(d)抽出
S121
TF(d)演算
S130
比較対象文書群P
読み出し
S140
索引語(P)抽出
S141
TF(P)演算
S142
IDF(P)演算
S150
類似度演算
S160
類似文書群S選出
S170
索引語(S)抽出
S171
IDF(S)演算
S180
特徴索引語抽出
エンド

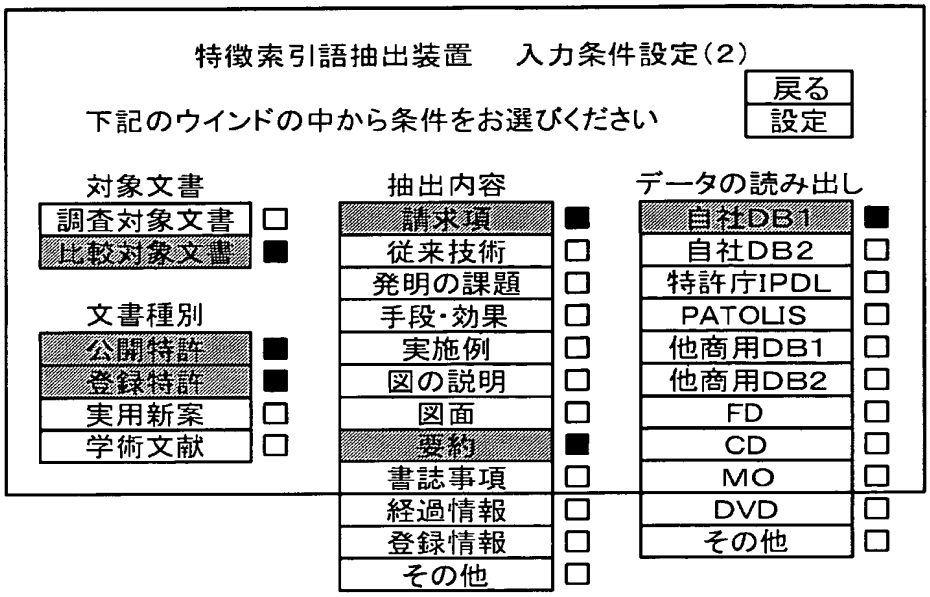
[図5]


出力処理スタート
S401
初期化
S402
出力条件の読み出し開始
S410
マップ作成条件読み出し
S420
リスト出力条件読み出し
S430
コメント追記条件読み出し
S411
マップ要
NO
YES
ストップ
S421
リスト要
NO
YES
ストップ
S431
コメント要
NO
YES
ストップ
S412
マップ用データ取り込み
S422
データ取り込み
S413
マップ生成
S423
リスト生成
S433
コメント生成
S440
マップ・リスト・コメント複合出力
エンド

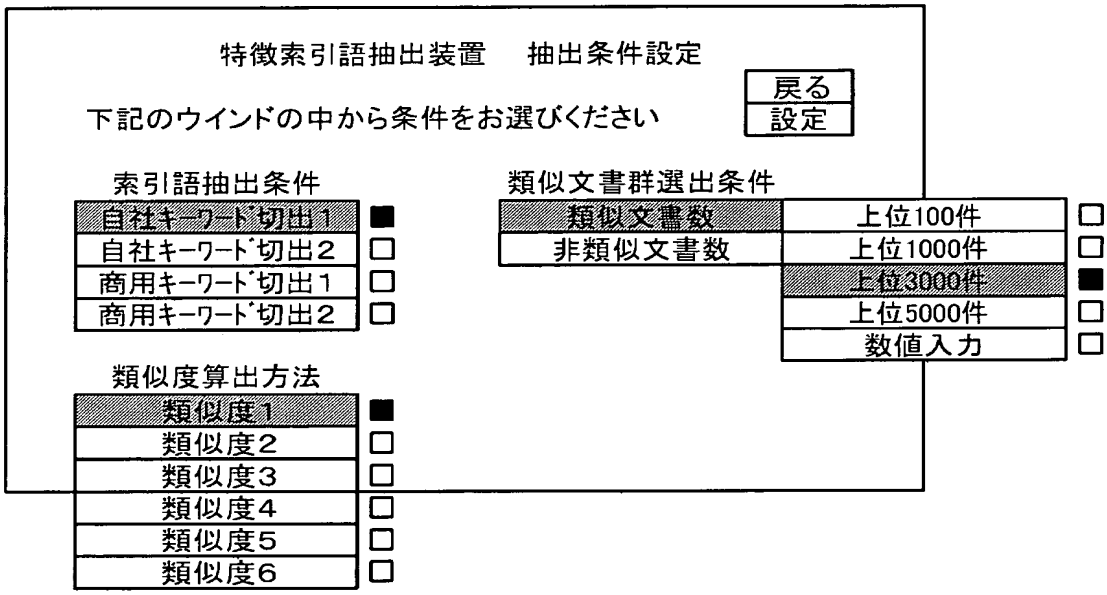
[図6]


特徴索引語抽出装置　入力条件設定(1)
下記のウインドの中から条件をお選びください
戻る
設定
対象文書
調査対象文書
比較対象文書
文書種別
公開特許
登録特許
実用新案
学術文献
データの読み出し
自社DB1
自社DB2
特許庁IPDL
PATOLIS
他商用DB1
他商用DB2
FD
CD
MO
DVD
その他
文書1
文書2
文書3
文書4
文書5
文書6

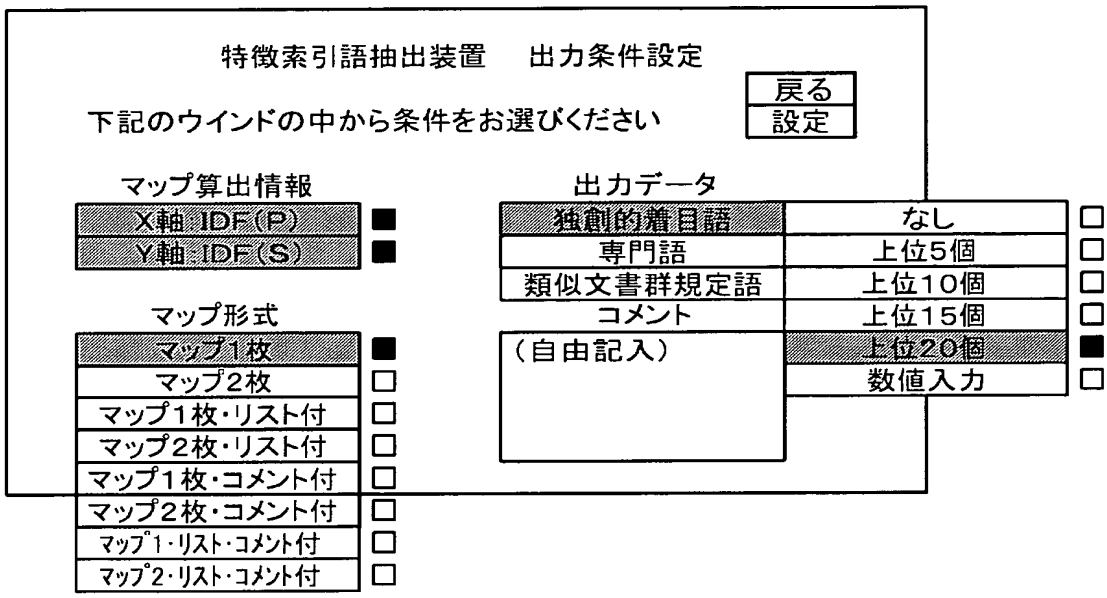
[図7]


特徴索引語抽出装置　入力条件設定(2)
戻る
設定
下記のウインドの中から条件をお選びください
対象文書
調査対象文書
比較対象文書
文書種別
公開特許
登録特許
実用新案
学術文献
抽出内容
請求項
従来技術
発明の課題
手段・効果
実施例
図の説明
図面
要約
書誌事項
経過情報
登録情報
その他
データの読み出し
自社DB1
自社DB2
特許庁IPDL
PATOLIS
他商用DB1
他商用DB2
FD
CD
MO
DVD
その他

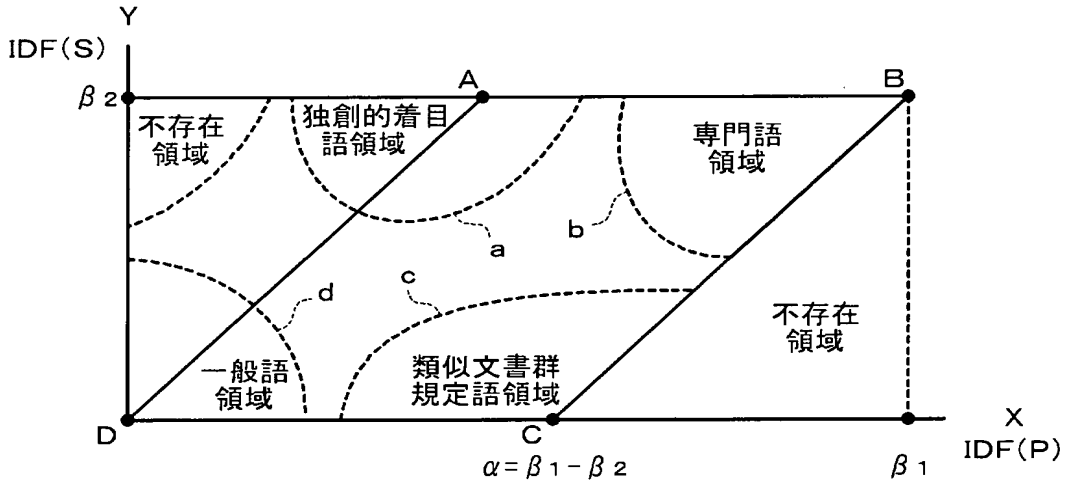
[図8]


特徴索引語抽出装置　抽出条件設定
戻る
設定
下記のウインドの中から条件をお選びください
索引語抽出条件
自社キーワード切出1
自社キーワード切出2
商用キーワード切出1
商用キーワード切出2
類似文書群選出条件
類似文書数
非類似文書数
上位100件
上位1000件
上位3000件
上位5000件
数値入力
類似度算出方法
類似度1
類似度2
類似度3
類似度4
類似度5
類似度6

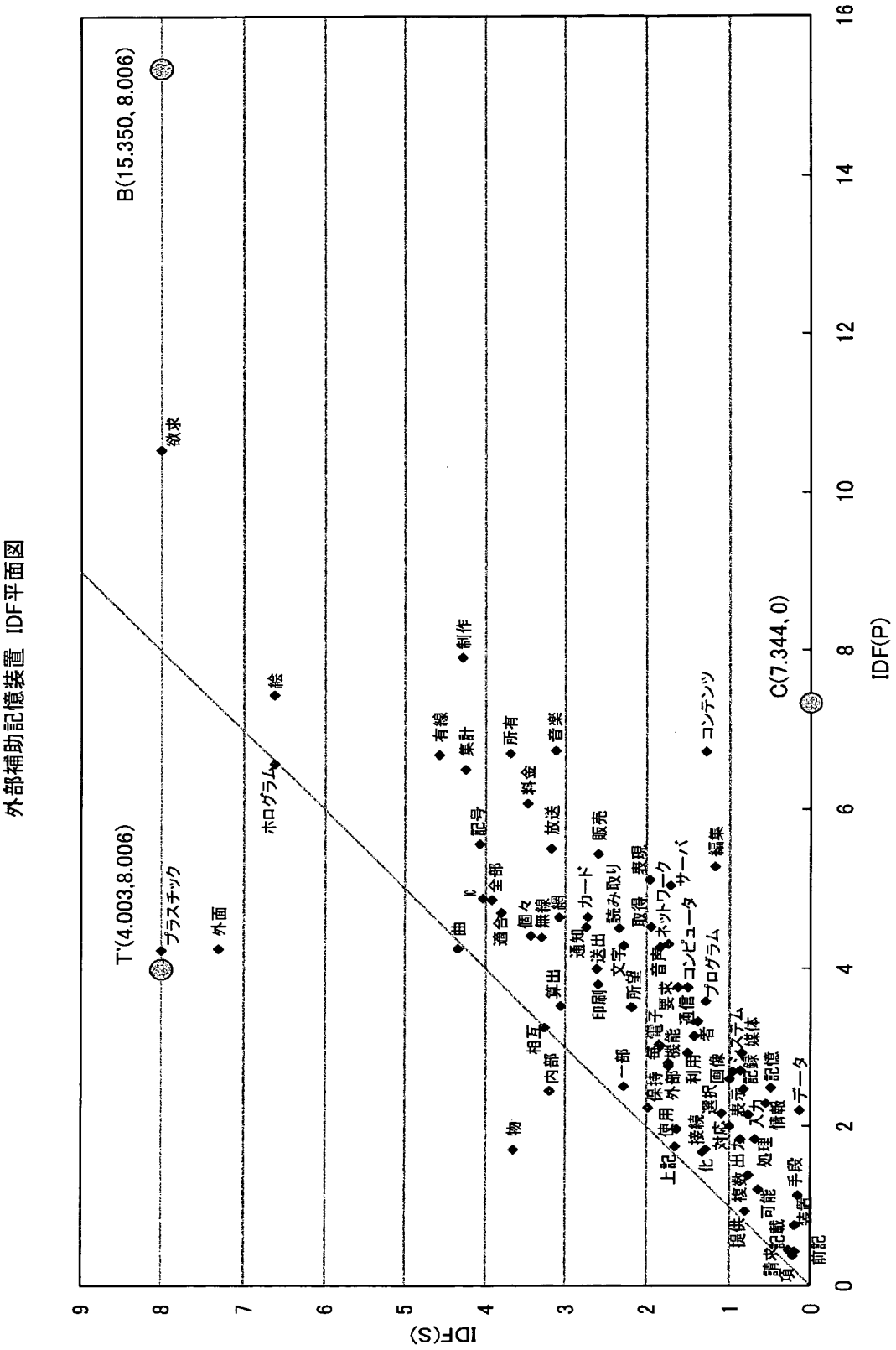
[図9]


特徴索引語抽出装置　出力条件設定
戻る
設定
下記のウインドの中から条件をお選びください
マップ算出情報
X軸 IDF(P)
Y軸 IDF(S)
マップ形式
マップ1枚
マップ2枚
マップ1枚・リスト付
マップ2枚・リスト付
マップ1枚・コメント付
マップ2枚・コメント付
ﾏｯﾌﾟ1・ﾘｽﾄ・ｺﾒﾝﾄ付
ﾏｯﾌﾟ2・ﾘｽﾄ・ｺﾒﾝﾄ付
出力データ
独創的着目語
専門語
類似文書群規定語
コメント
（自由記入）
なし
上位5個
上位10個
上位15個
上位20個
数値入力

[図10]


Y
IDF(S)
β2
A
B
不存在
領域
独創的着目
語領域
専門語
領域
a
b
c
d
一般語
領域
類似文書群
規定語領域
不存在
領域
D
C
X
IDF(P)
α=β1−β2
β1

[図11]

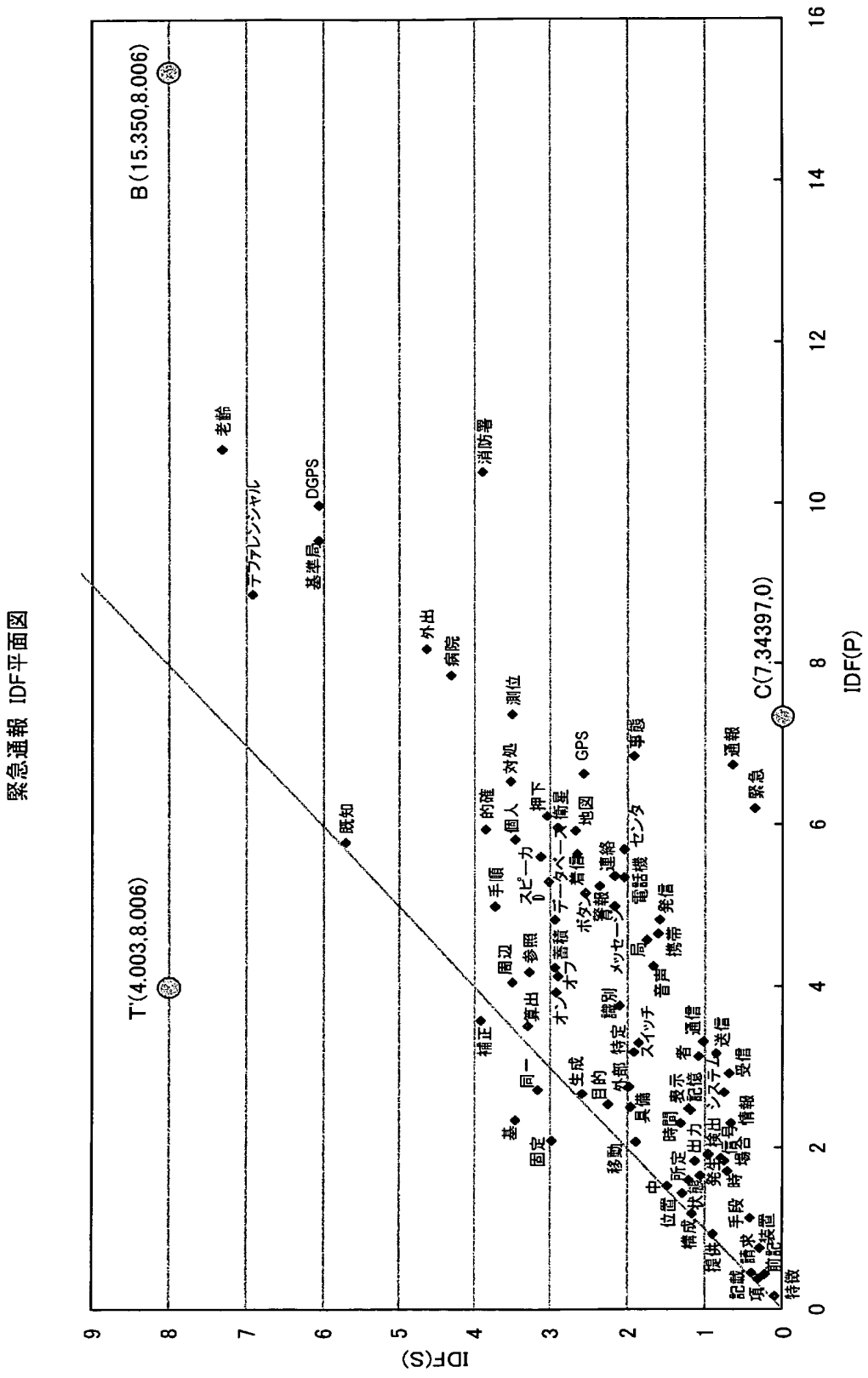

外部補助記憶装置 IDF平面図
B(15.350, 8.006)
T'(4.003,8.006)
C(7.344, 0)
IDF(S)
IDF(P)
欲求
プラスチック
外面
ホログラム
絵
有線
集計
制作
所有
料金
放送
音楽
販売
カード
通知
送出
読み取り
取得
表現
ネットワーク
サーバ
コンピュータ
プログラム
コンテンツ
編集
記号
全部
曲
適合
個々
無線
網
算出
相互
内部
物
印刷
文字
所望
一部
保持
電子
要求
音声
機能
通信
者
利用
外部
使用
上記
接続
化
対応
選択
画像
表示
入力
記録
媒体
システム
記憶
情報
データ
処理
可能
手段
装置
複数
出力
提供
記載
請求項
前記

[図12]

外部補助記憶装置　各領域から20個まで抽出

X'値順	a:独創的着目語	b:専門語	c:類似文書群規定語	d:一般語
1	絵		コンテンツ	上記
2	ホログラム		編集	提供
3	記号		表現	請求
4	IC		サーバ	項
5	全部		取得	記載
6	適合		ネットワーク	前記
7	網		音声	保持
8	カード		コンピュータ	使用
9	通知		要求	化
10	個々		プログラム	接続
11	無線		通信	可能
12	文字		者	装置
13	外面		毎	複数
14	曲		電子	手段
15	プラスチック		媒体	出力
16	送出		利用	対応
17	印刷		機能	外部
18	算出		外部	機能
19	所望		記録	選択
20	通信		システム	処理

[図13]

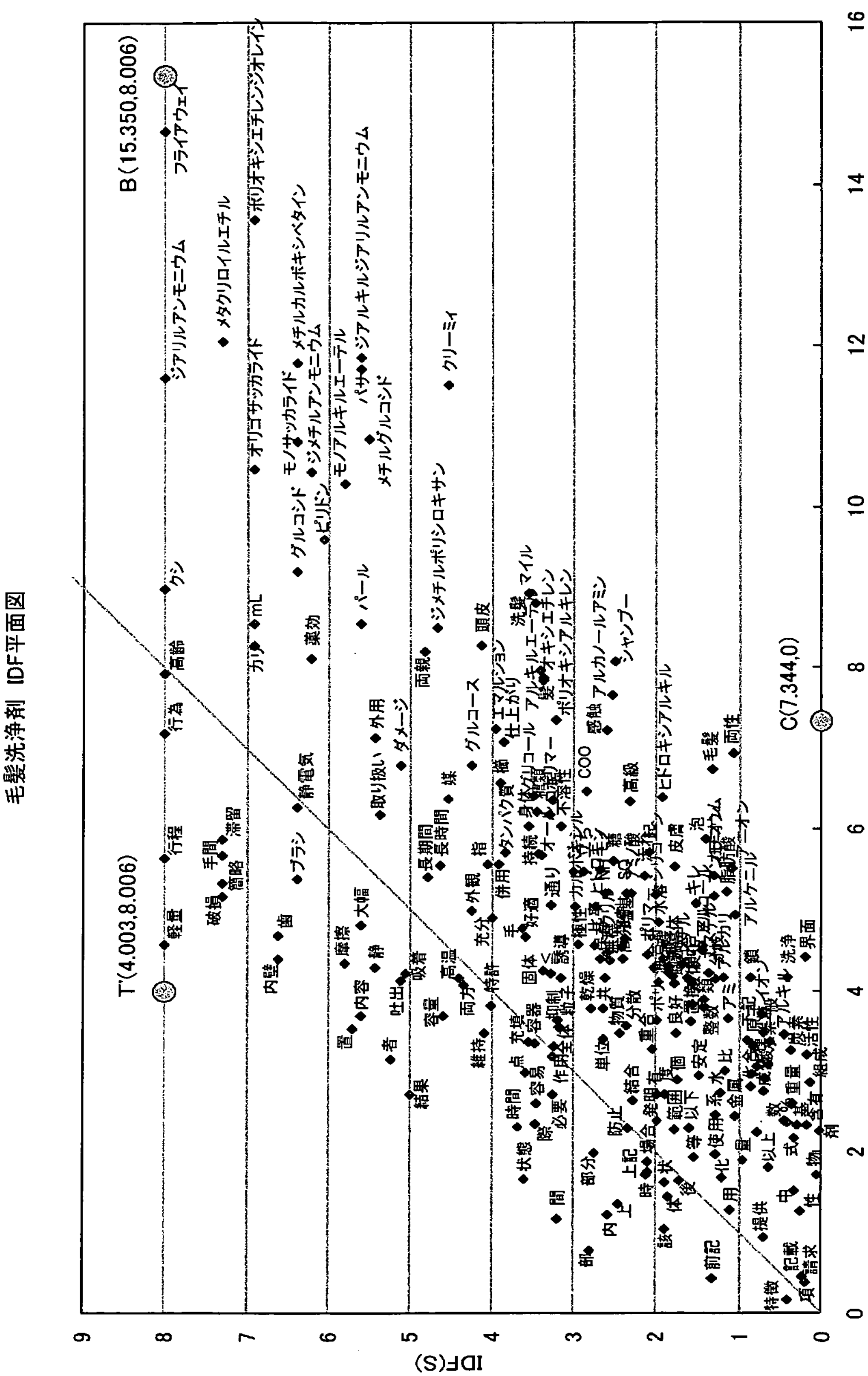

緊急通報 IDF平面図
T'(4.003,8.006)
B(15.350,8.006)
C(7.34397,0)
IDF(S)
IDF(P)
老齢
デファレンシャル
DGPS
基準局
消防署
外出
病院
測位
既知
的確
個人
対処
押下
衛星
地図
GPS
事態
センタ
電話機
連絡
着信
データベース
蓄積
オプ
ボタン
音声
携帯
発信
局
メッセージ
着信
識別
待定
スイッチ
外部
目的
生成
同一
算出
参照
周辺
補正
基
固定
移動
具備
中
所定
位置
状態
構成
提供
記載
項
請求
手段
装置
時
場合
出力
時間
表示
記憶
システム
通信
送信
受信
情報
発
通報
緊急
待機

[図14]

緊急通報　各領域から20個まで抽出

X'値順	a:独創的着目語	b:専門語	c:類似文書群規定語	d:一般語
1	デファレンシャル	消防署	事態	構成
2	既知		通報	中
3	手順		緊急	特徴
4	データベース		センタ	提供
5	蓄積		電話機	請求
6	参照		発信	項
7	オフ		携帯	記載
8	周辺		局	位置
9	オン		音声	移動
10	識別		通信	前記
11	補正		スイッチ	所定
12	算出		特定	装置
13	スイッチ		送信	具備
14	特定		者	状態
15	者		受信	手段
16	外部		外部	出力
17	同一		システム	外部
18	システム		具備	検出
19	生成		記憶	発生
20	目的		表示	時間

[図15]

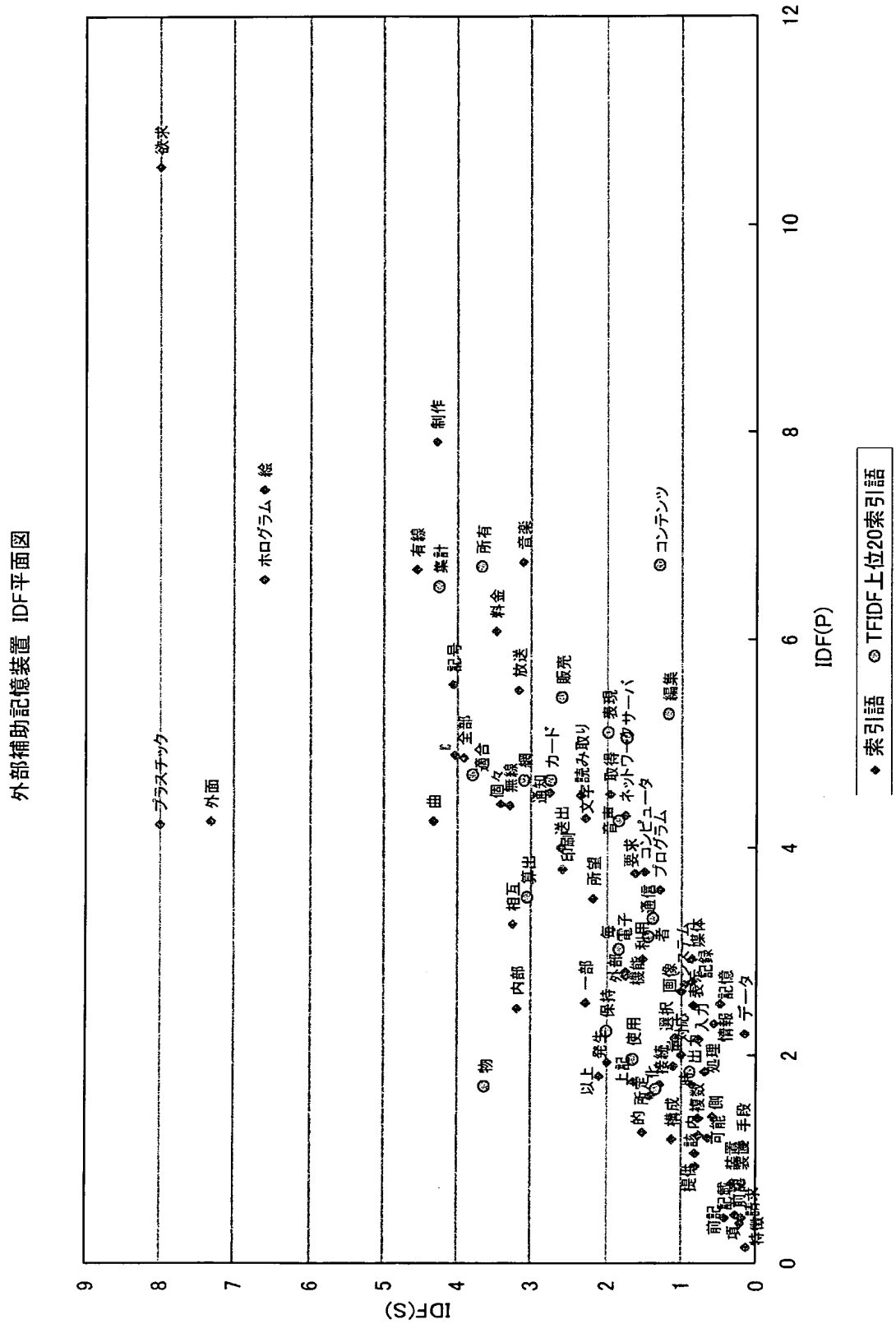

毛髪洗浄剤 IDF平面図
B(15.350,8.006)
T'(4.003,8.006)
C(7.344,0)
IDF(P)
IDF(S)
フライアウェイ
ポリオキシエチレンジオレイン
ジアリルアンモニウム
メタクリロイルエチル
メチルカルボキシベタイン
ジアルキルジアリルアンモニウム
クリーミィ
ジメチルポリシロキサン
シャンプー
アルカノールアミン
ヒドロキシアルキル
アルケニル

[図16]

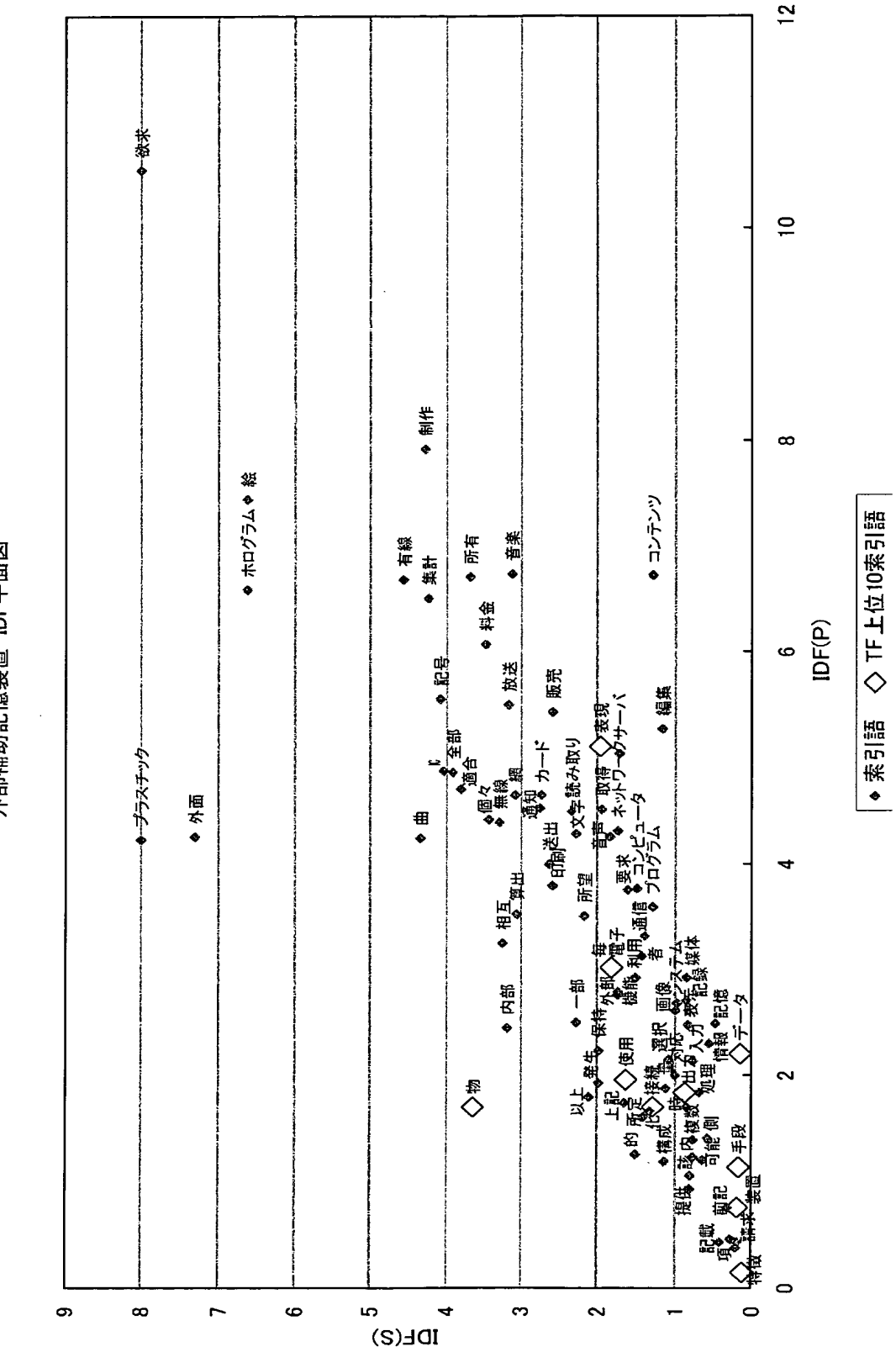
毛髪洗浄剤　各領域から20個まで抽出

X'値順	a:独創的着目語	b:専門語	c:類似文書群規定語	d:一般語
1	クシ	フライアウェイ	両性	前記
2	mL	ポリオキシエチレンジオレイン	毛髪	該
3	カリ	メチルカルボキシベタイン	ヒドロキシアルキル	体
4	薬効	ジアルキルジアリルアンモニウム	泡	状
5	高齢	パサ	皮膚	特徴
6	行為	メチルグルコシド	アニオン	後
7	外用	クリーミィ	カチオン	用
8	ダメージ	マイル	脂肪酸	請求
9	媒	洗髪	シリコーン	項
10	静電気	シャンプー	アンモニウム	記載
11	取り扱い		キレ	提供
12	滞留		アルケニル	発明
13	タンパク質		水溶	等
14	手間		アル	化
15	行程		アルコール	範囲
16	併用		界面	使用
17	指		残	以下
18	長時間		メチル	有
19	長期間		感	度
20	ブラシ		合計	量

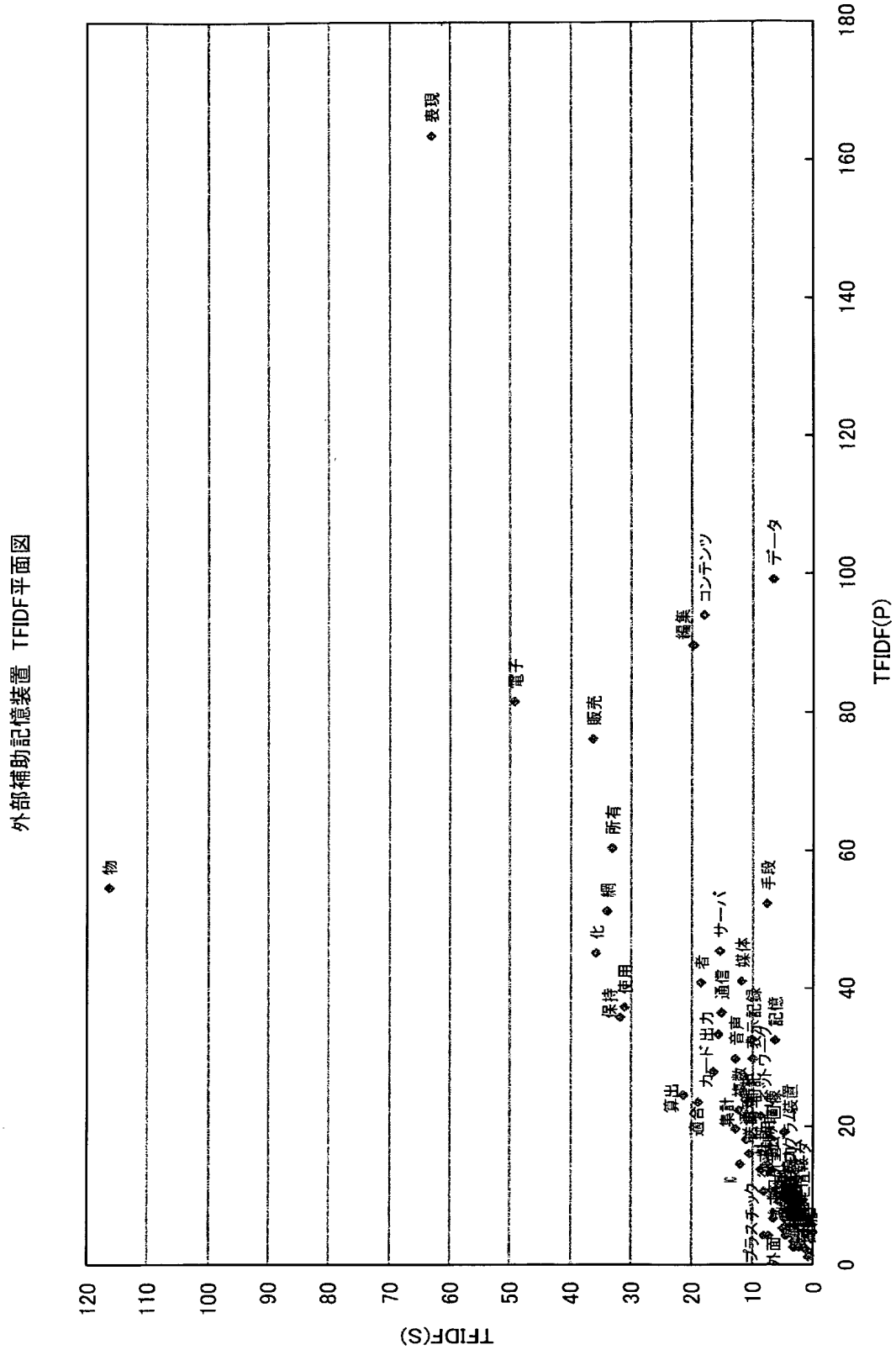
[図17]


外部補助記憶装置 IDF平面図
IDF(S)
IDF(P)
◆索引語 ◎TFIDF上位20索引語
プラスチック
欲求
外面
ホログラム
絵
有線
集計
所有
制作
曲
記号
料金
音楽
放送
販売
カード
印刷送出
文字読み取り
所望
音声
取得
表現
ネットワーク
サーバ
要求
コンピュータ
プログラム
コンテンツ
編集
物
内部
相互
算出
一部
保持
外部
電子
使用
上記
以上
的
構成
選択
画像
入力
記録
媒体
データ
情報
処理
出力
装置
手段
可能
該
複数
前記

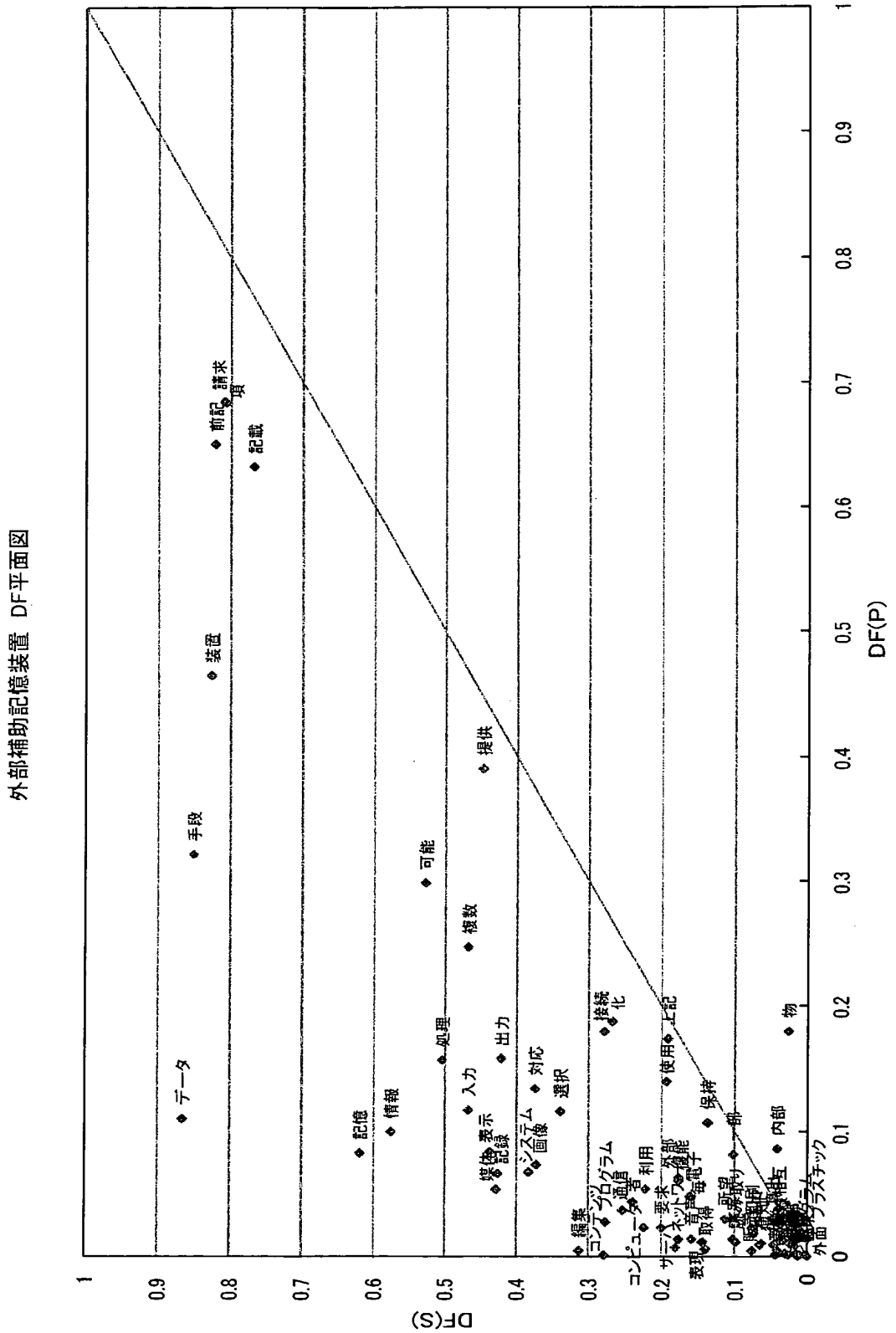
[図18]


外部補助記憶装置 IDF平面図
IDF(S)
IDF(P)
◆索引語 ◇TF上位10索引語
プラスチック
欲求
外面
ホログラム
絵
有線
集計
制作
所有
曲
記号
料金
音楽
全部
適合
個々
無線
網
放送
販売
カード
通知
送出
所望
相互
一部
内部
物
保持
外部
電子
機能
利用
通信
要求
コンピュータ
ネットワーク
サーバ
取得
表現
音声
プログラム
コンテンツ
編集
以上
発生
使用
上記
的
承認
接続
構成
選択
画像
システム
媒体
記録
データ
処理
入力
出力
該
複数
可能
側
手段
複体
記載
請求
装置
特徴
項

[図19]

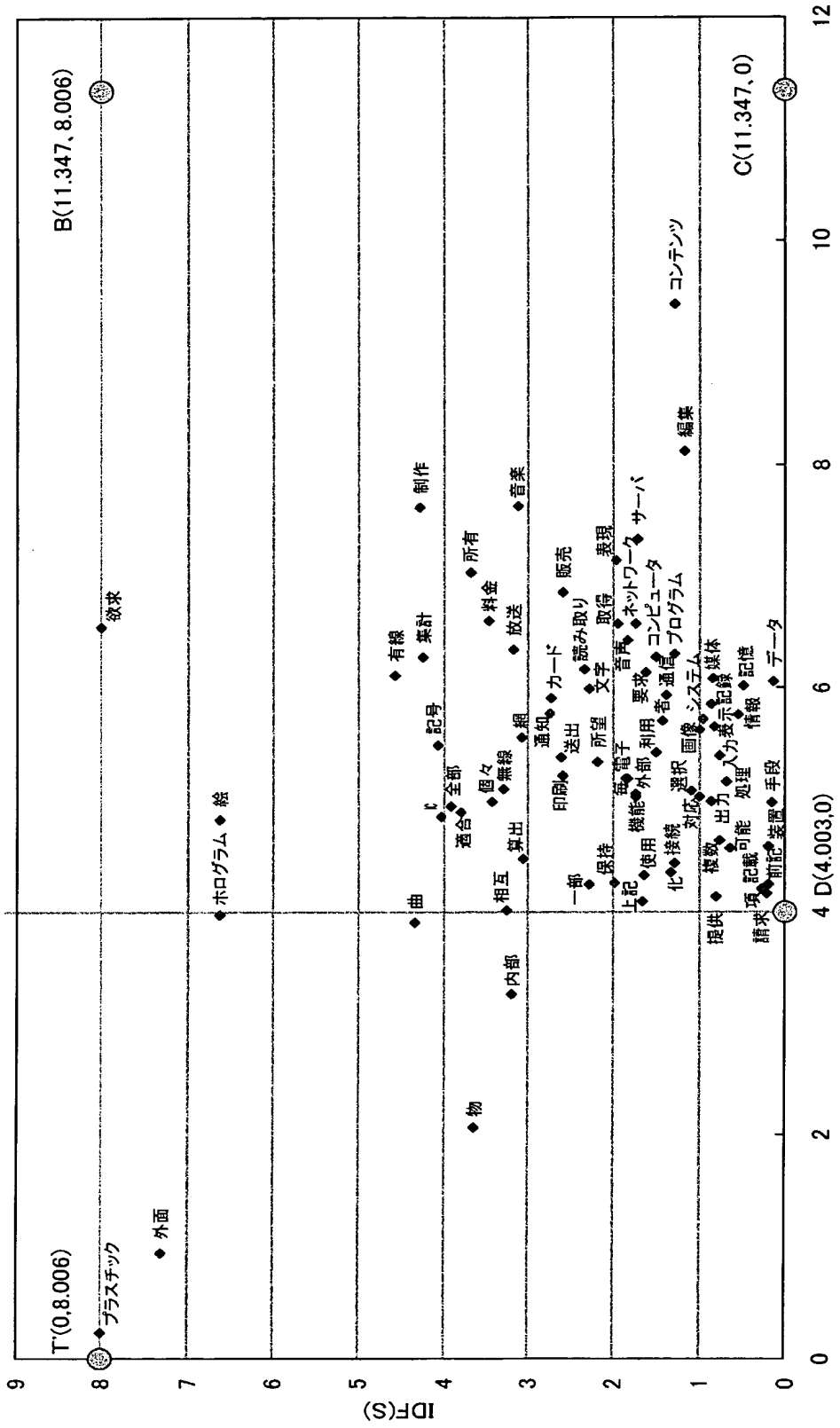

外部補助記憶装置 TFIDF平面図
TFIDF(S)
TFIDF(P)
物
表現
電子
販売
編集
コンテンツ
データ
所有
網
化
保持
使用
手段
媒体
サーバ
通信
音声
カード出力
算出
適合
記憶
0
10
20
30
40
50
60
70
80
90
100
110
120
0
20
40
60
80
100
120
140
160
180

[図20]

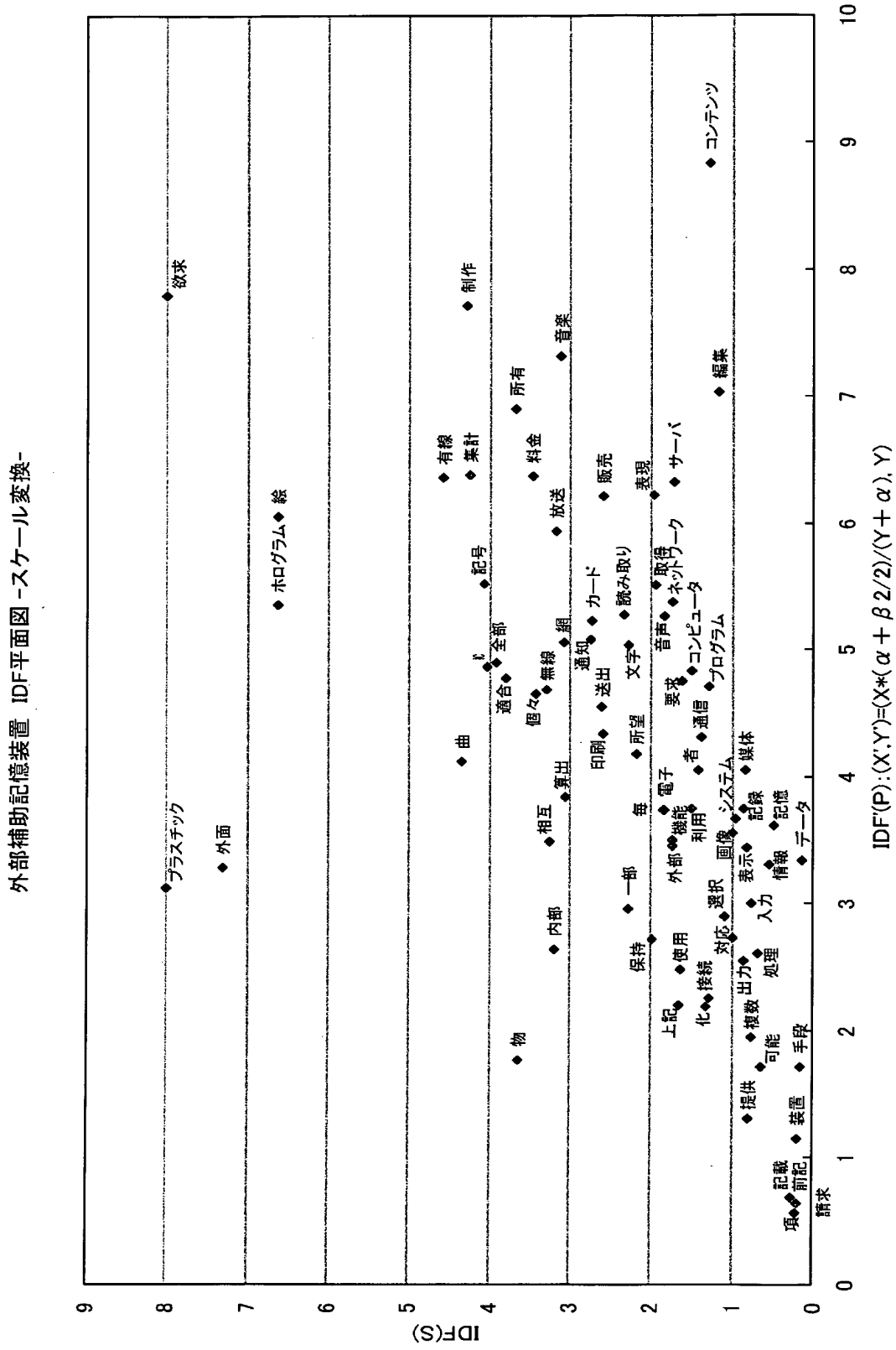

外部補助記憶装置 DF平面図
DF(S)
DF(P)
データ
手段
装置
前記
請求項
記載
可能
複数
提供
処理
入力
出力
対応
選択
記憶
情報
表示
記録
システム
画像
接続
化
使用
保持
内部
物
外面プラスチック
コンピュータ
通信
利用
コンピュータプログラム
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

[図21]

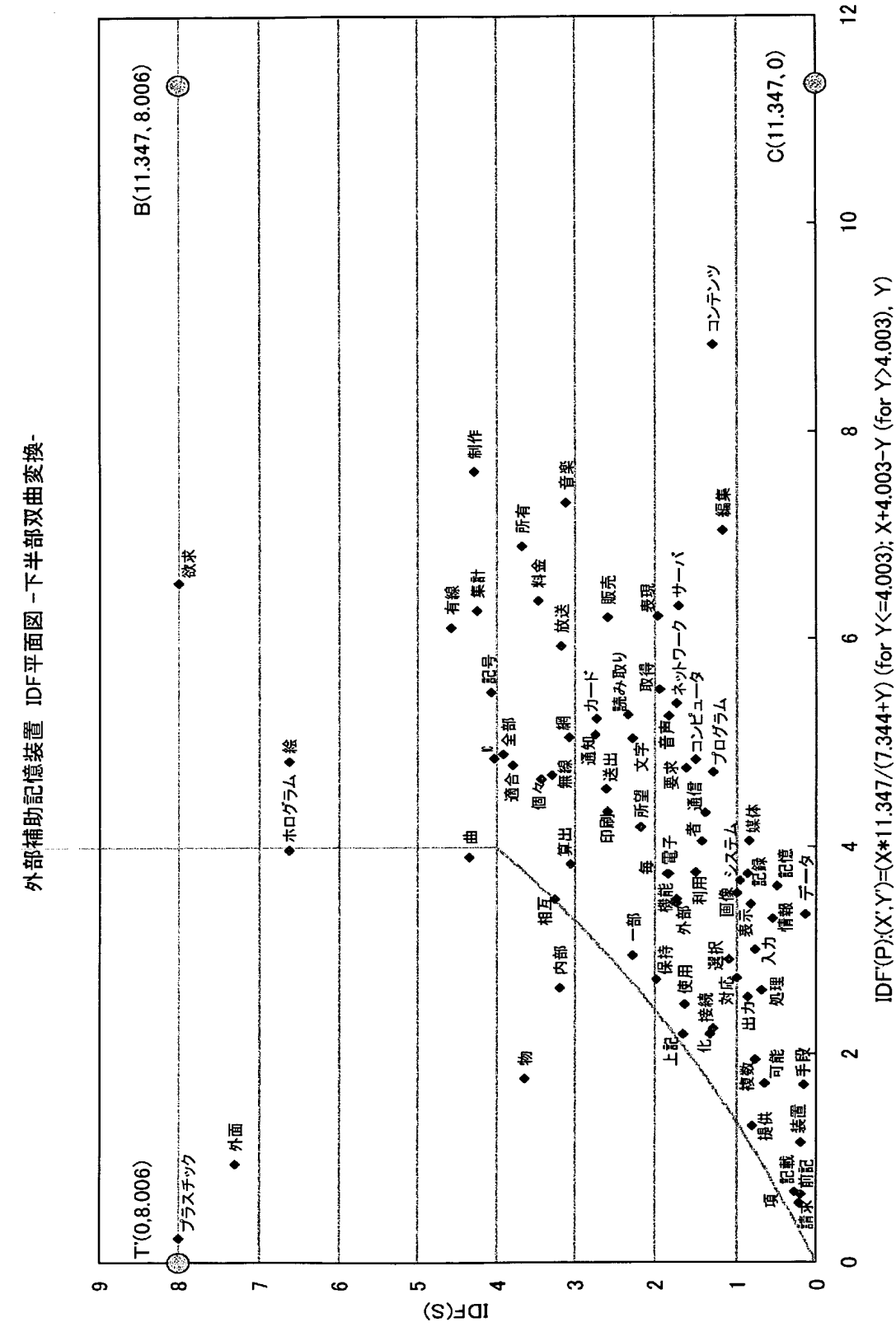
外部補助記憶装置 IDF平面図 -線形変換-



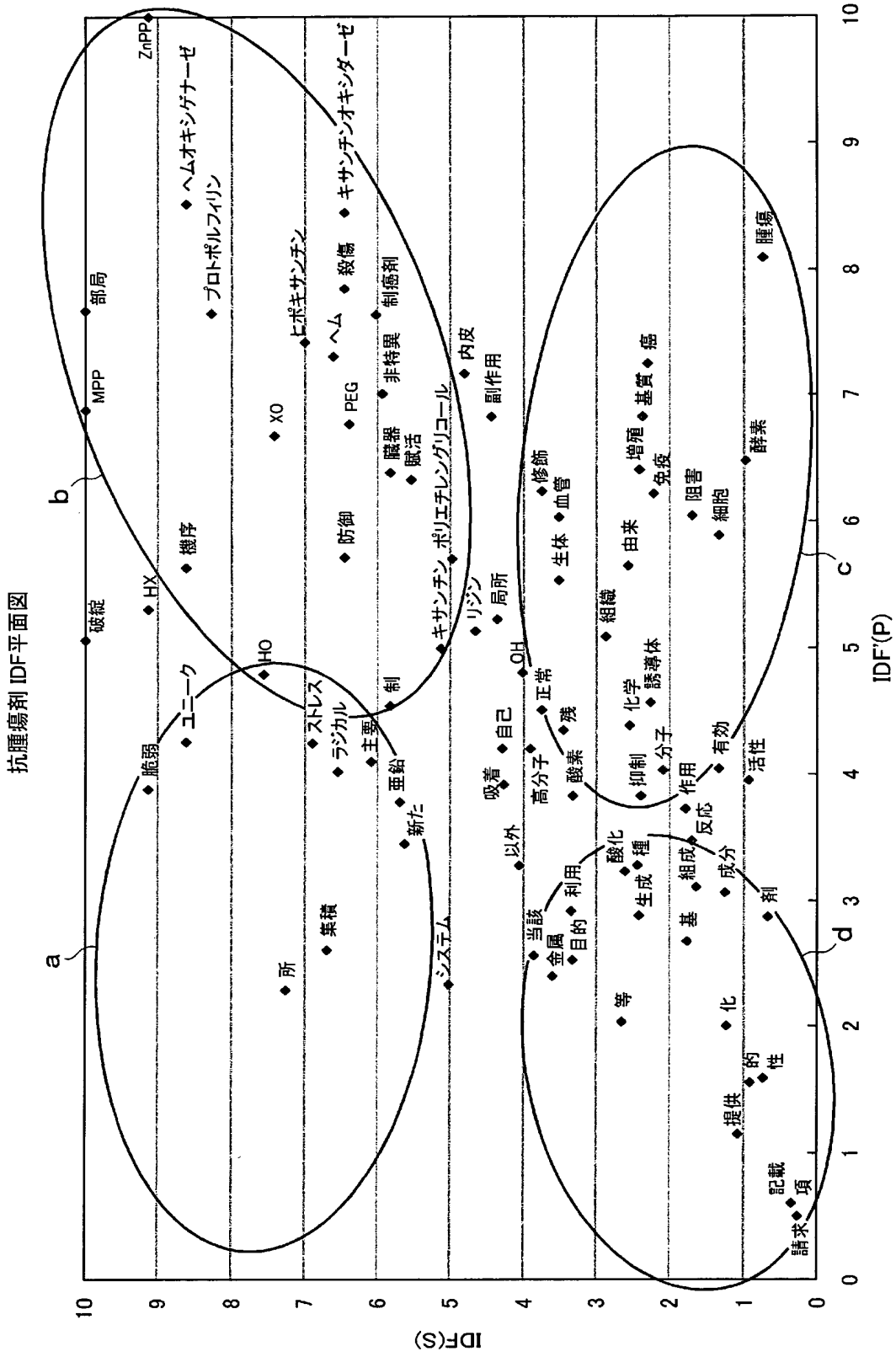
[図22]


外部補助記憶装置 IDF平面図 -スケール変換-
IDF(S)
IDF'(P):(X',Y')=(X*(α+β2/2)/(Y+α), Y)
0
1
2
3
4
5
6
7
8
9
10
請求
記載
前記
装置
手段
提供
可能
複数
出力
処理
入力
情報
記憶
データ
表示
システム
画像
選択
対応
接続
化
上記
保持
使用
内部
一部
毎
相互
算出
外部
機能
利用
電子
所望
曲
者
通信
プログラム
適合
個々
無線
通知
送出
印刷
カード
読み取り
文字
要求
音声
コンピュータ
ネットワーク
取得
表現
販売
サーバ
コンテンツ
編集
音楽
放送
料金
所有
有線
集計
制作
記号
全部
ホログラム
絵
外面
プラスチック
欲求

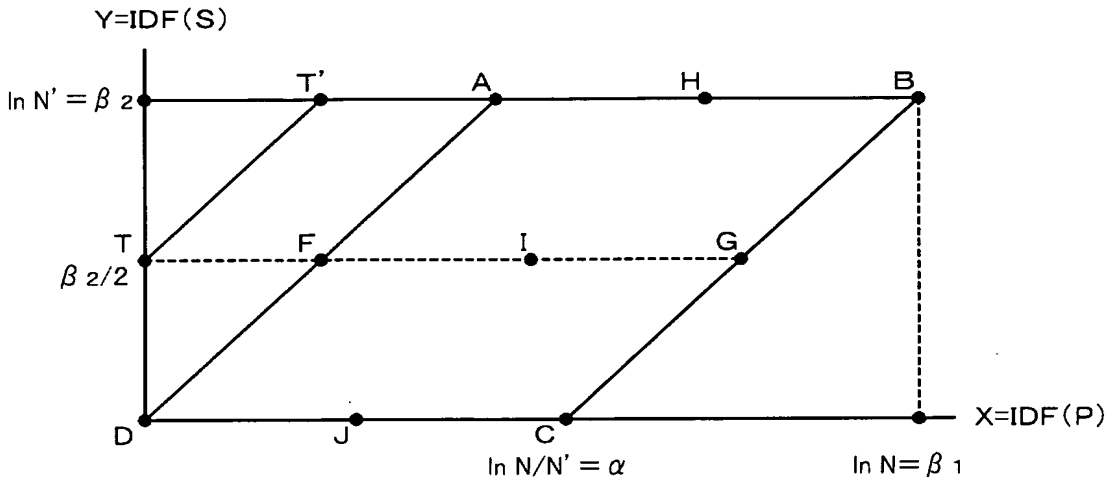
[図23]


外部補助記憶装置 IDF平面図 -下半部双曲変換-
T'(0,8.006)
B(11.347,8.006)
C(11.347,0)
IDF(S)
IDF'(P):(X',Y')=(X*11.347/(7.344+Y)) (for Y<=4.003); X+4.003-Y (for Y>4.003), Y)
プラスチック
外面
欲求
ホログラム
絵
曲
有線
集計
制作
記号
所有
料金
音楽
全部
適合
個々
無線
網
放送
販売
通知
カード
読み取り
印刷
送出
取得
表現
算出
相互
所望
文字
要求
音声
ネットワーク
サーバ
コンピュータ
コンテンツ
プログラム
編集
一部
毎
電子
機能
通信
内部
利用
者
媒体
システム
記録
物
保持
外部
使用
表示
記憶
上記
接続
選択
入力
情報
化
対応
出力
処理
複数
可能
提供
手段
装置
記載
前記
項
請求
データ

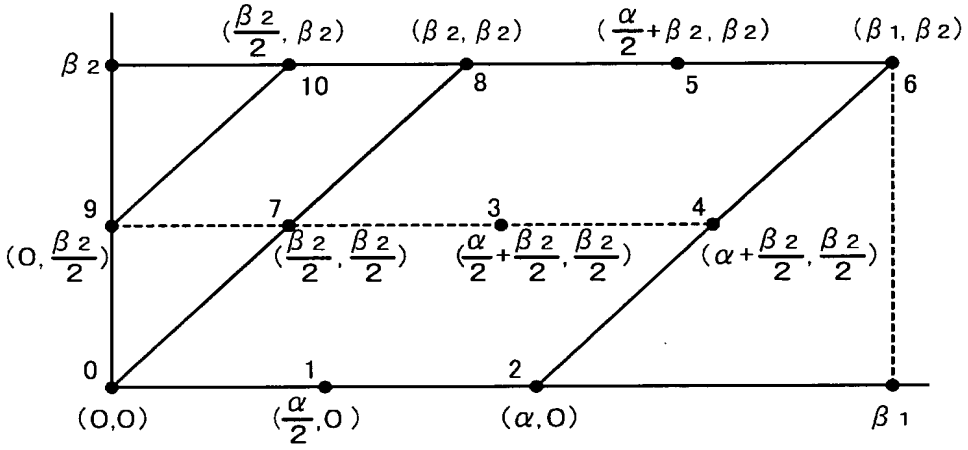
[図24]


抗腫瘍剤 IDF平面図
IDF(S)
IDF'(P)
a
b
c
d
ZnPP
部局
MPP
ヘムオキシゲナーゼ
プロトポルフィリン
ヒポキサンチン
ヘム
殺傷
キサンチンオキシダーゼ
制癌剤
内皮
副作用
非特異
PEG
XO
臓器
賦活
HX
機序
破綻
HO
防御
キサンチン
ポリエチレングリコール
リジン
局所
修飾
血管
生体
増殖
基質
癌
免疫
由来
組織
阻害
細胞
酵素
腫瘍
脆弱
ユニーク
ストレス
ラジカル
主要
制
亜鉛
新た
吸着
自己
高分子
以外
正常
残
酸素
化学
誘導体
分子
抑制
作用
有効
活性
酸化
種
生成
組成
反応
成分
剤
利用
目的
当該
金属
システム
所
集積
等
基
化
提供
的
性
記載
項
請求

[図25]

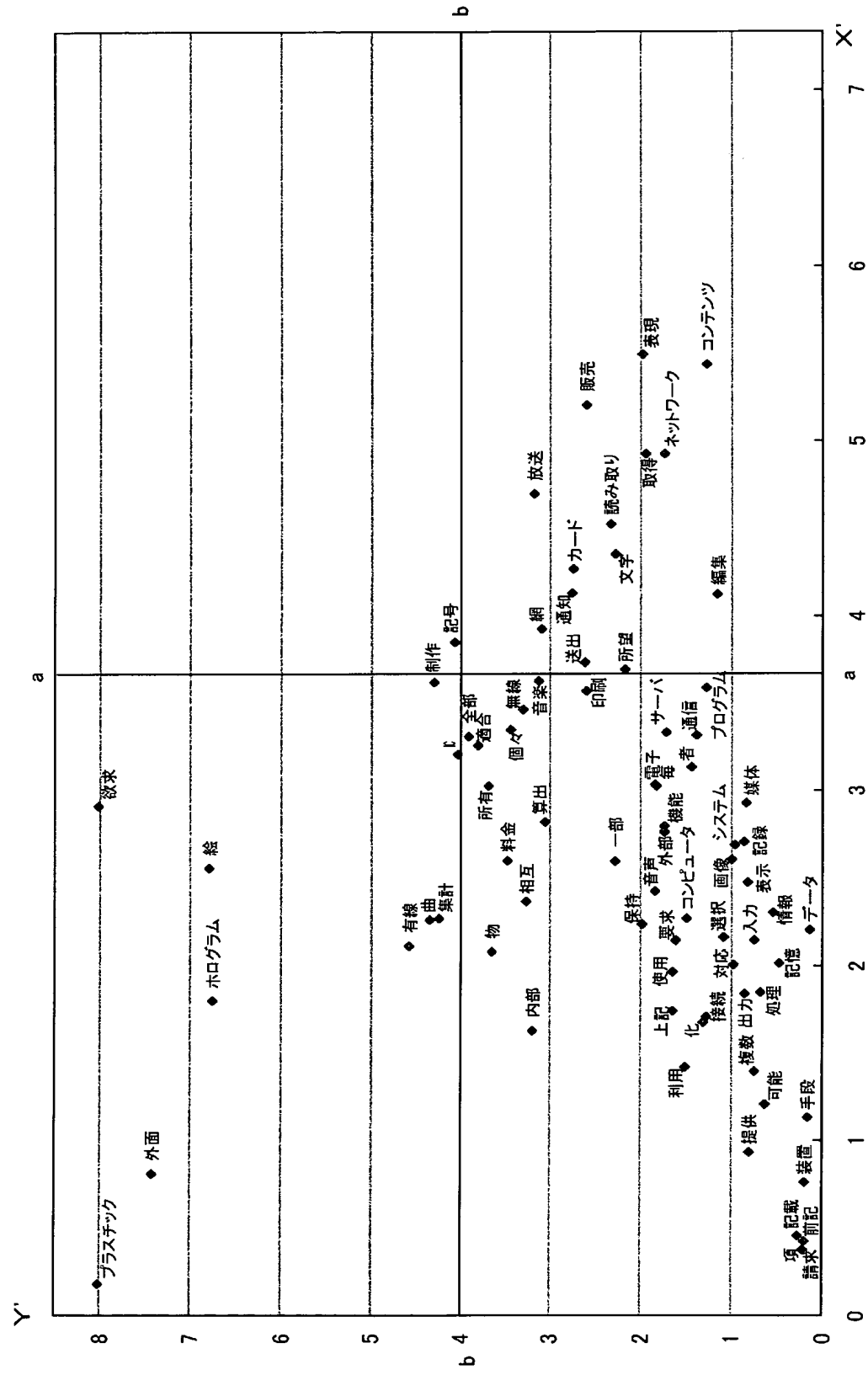


[図26]

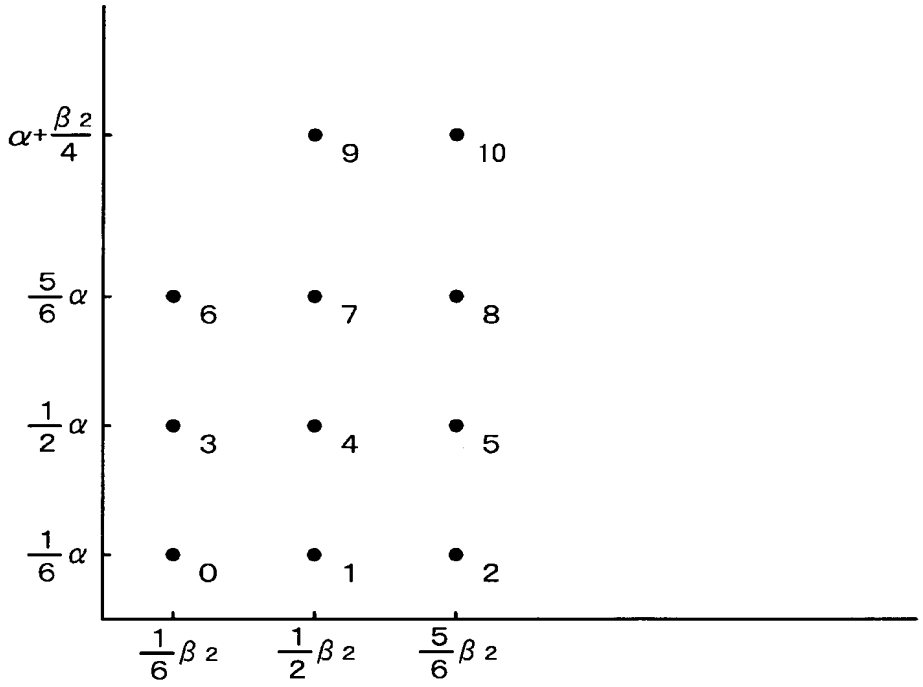


SOM応用例1の参照点

[図27]


外部補助記憶装置 SOM応用例1 -11点斜方格子-
Y'
X'
a
b
プラスチック
外面
欲求
ホログラム
給
有線
曲
集計
物
内部
所有
料金
相互
算出
制作
記号
全部
適合
個々
無線
音楽
一部
音声
外部
電子
サーバ
機能
者
通信
プログラム
コンピュータ
選択
画像
システム
媒体
入力
表示
記録
情報
データ
保持
要求
使用
上記
化
接続
対応
出力
処理
記憶
複数
利用
可能
提供
手段
装置
記載
前記
請求
項
送出
通知
印刷
所望
網
放送
カード
文字
読み取り
販売
取得
ネットワーク
表現
コンテンツ
編集

[図28]

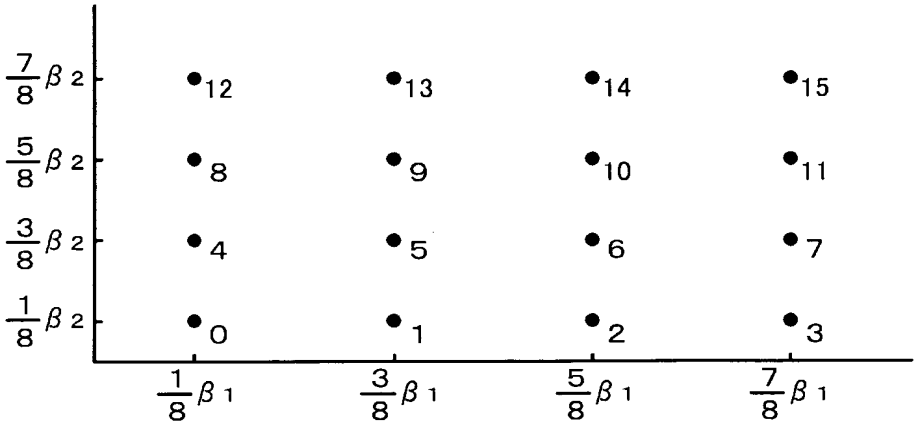


SOM応用例2の参照点

[図29]

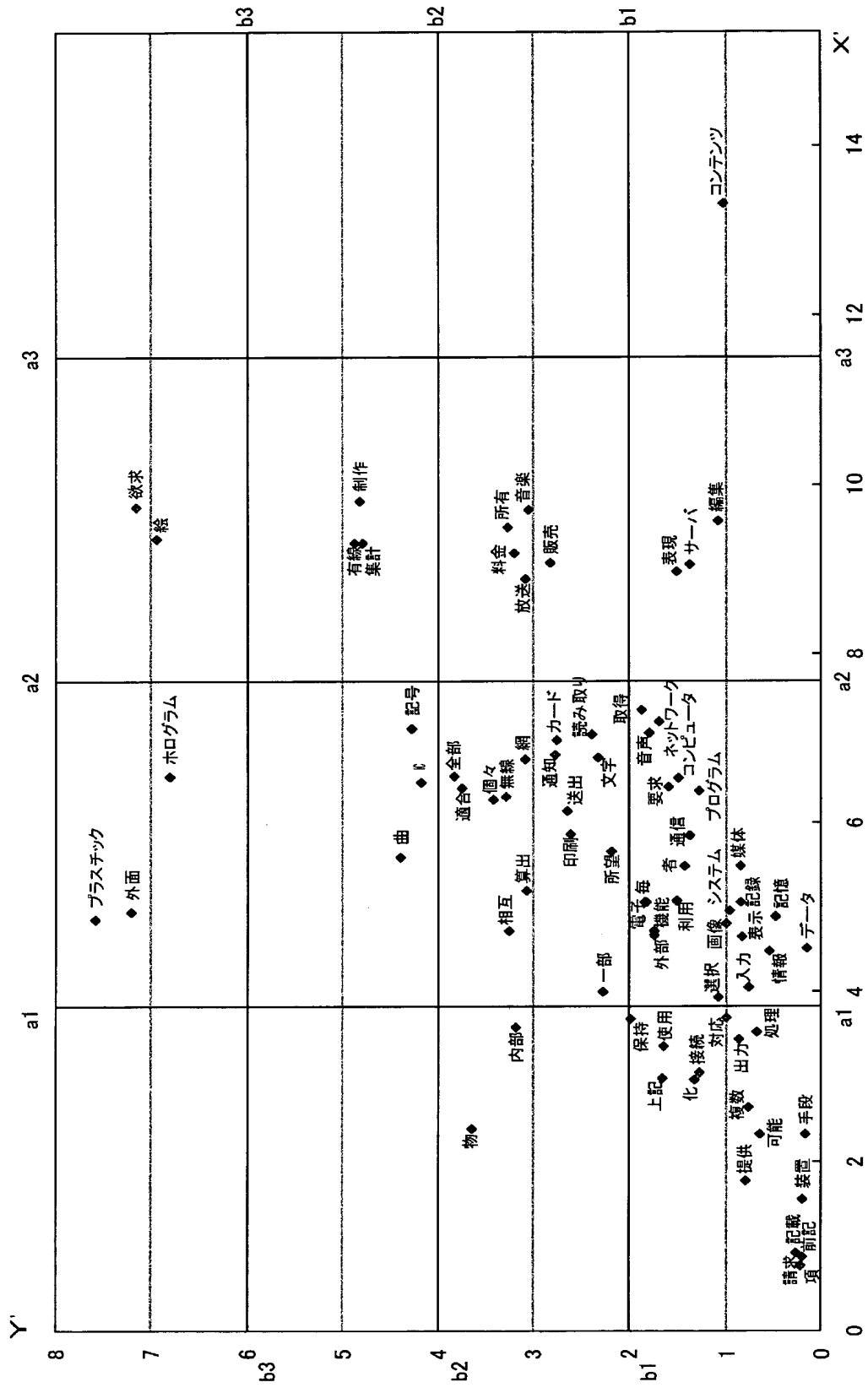

外部補助記憶装置 SOM応用例2 -11点格子逆変換-
Y'
X'
a
b
プラスチック
外面
ホログラム
絵
欲求
曲
ℓ
有線
集計
記号
創作
所有
料金
放送
音楽
販売
表現
サーバ
取得
ネットワーク
編集
コンテンツ
システム
記録
表示
情報
画像
利用
者
選択
入力
処理
対応
出力
可能
手段
装置
複数
提供
媒体
記憶
データ
記載
前記
項
請求
化
接続
使用
上記
保持
プログラム
コンピュータ
音声
要求
通信
外部
機能
電子
所望
送出
印刷
通知
カード
読み取り
文字
一部
算出
網
無線
物
全部
適合
個々
内部
相互

[図30]



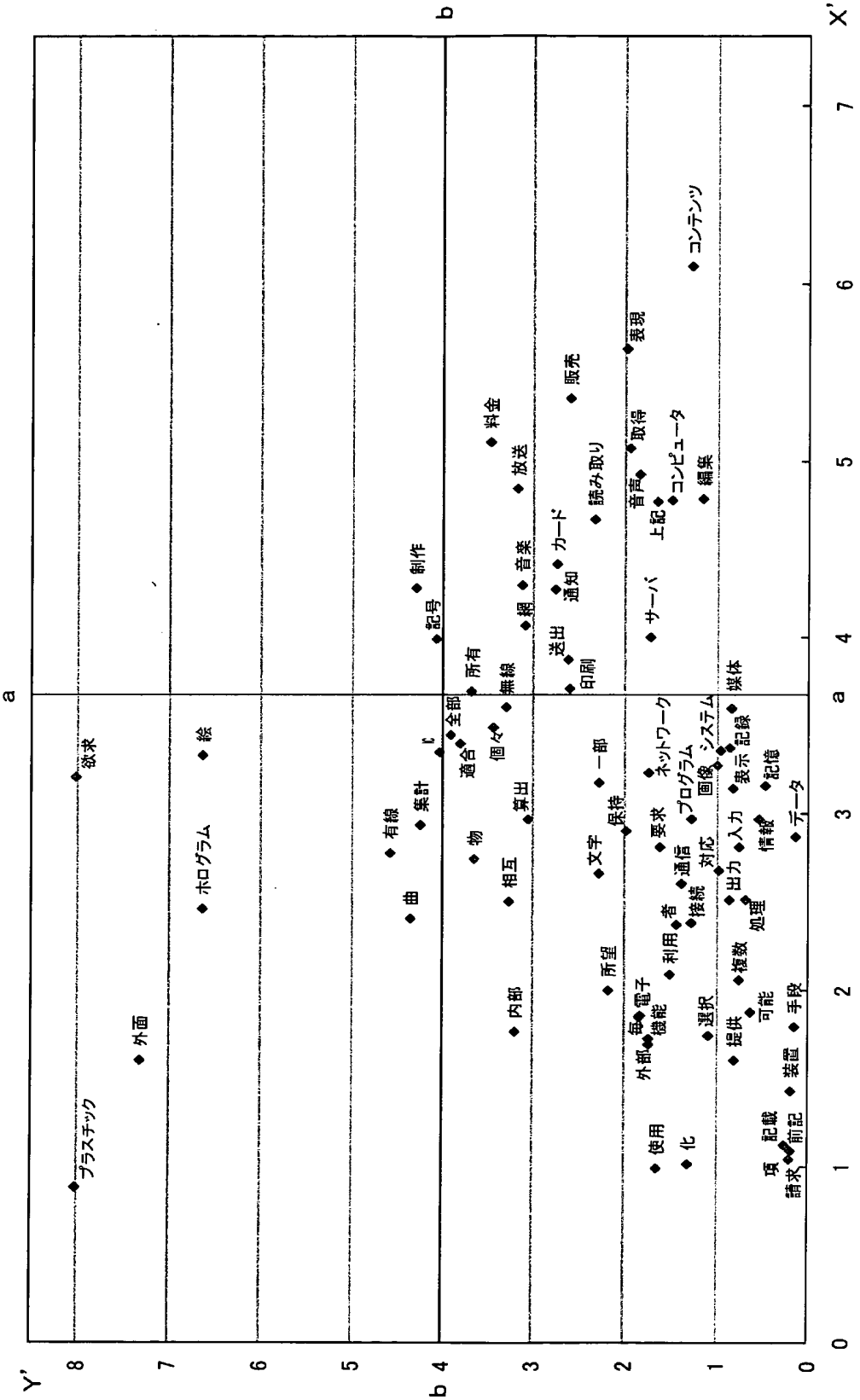
SOM応用例3の参照点

[図31]

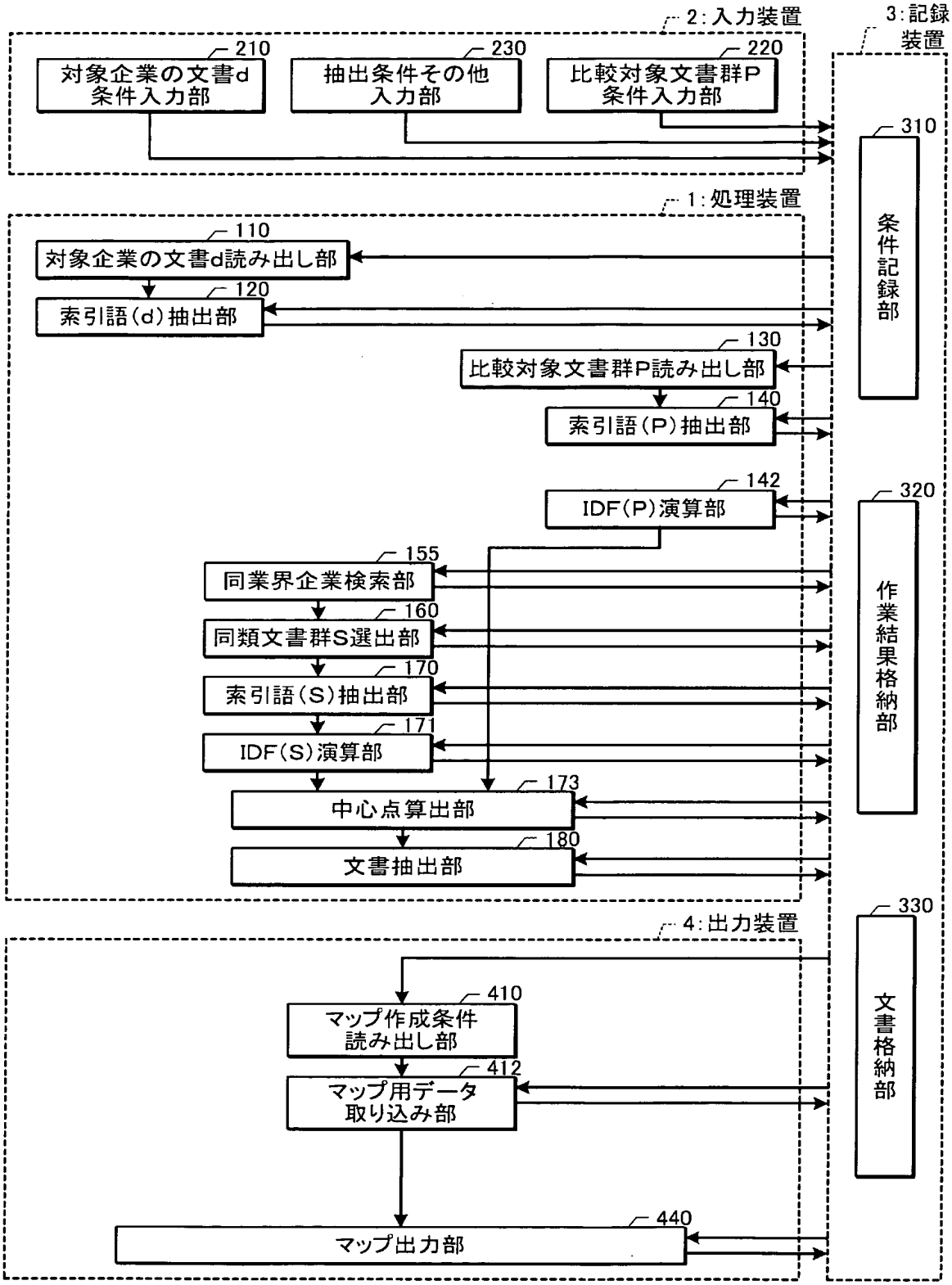

外部補助記憶装置 SOM応用例3 -16点スケール変換-
Y'
X'
a1
a2
a3
b1
b2
b3
プラスチック
外面
ホログラム
欲求
絵
曲
記号
有線
集計
制作
適合
全部
個々
無線
網
相互
算出
内部
料金
放送
所有
音楽
販売
通知
カード
読み取り
送出
印刷
所望
文字
取得
一部
物
保持
使用
上記
化
接続
対応
出力
複数
可能
提供
処理
手段
装置
記載
前記
請求項
電子
毎
外部
機能
利用
音声
ネットワーク
コンピュータ
要求
表現
サーバ
編集
コンテンツ
着
通信
プログラム
システム
媒体
選択
画像
入力
表示
記録
情報
記憶
データ

[図32]

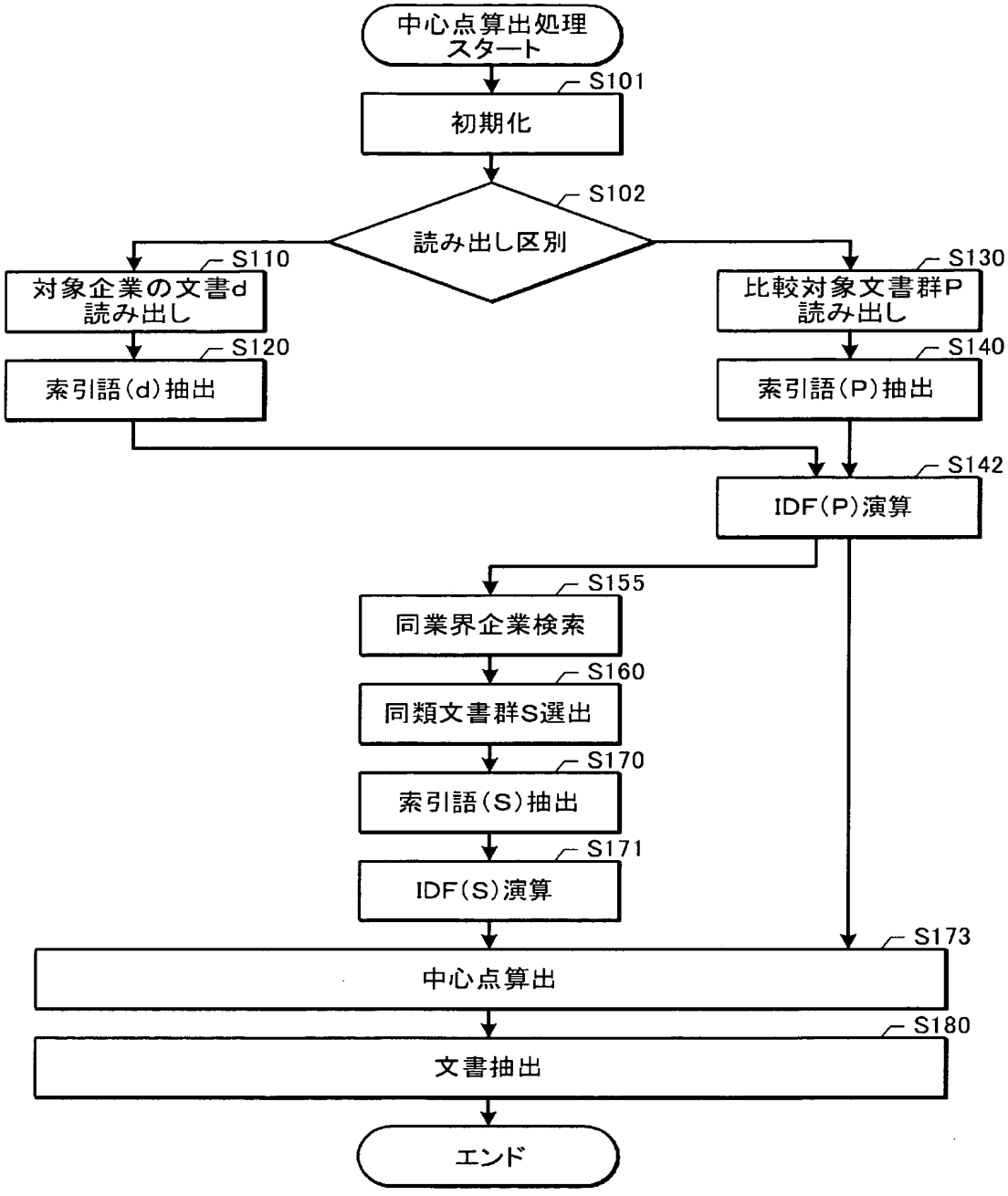
外部補助記憶装置　SOM 応用例4 -多次元入力-



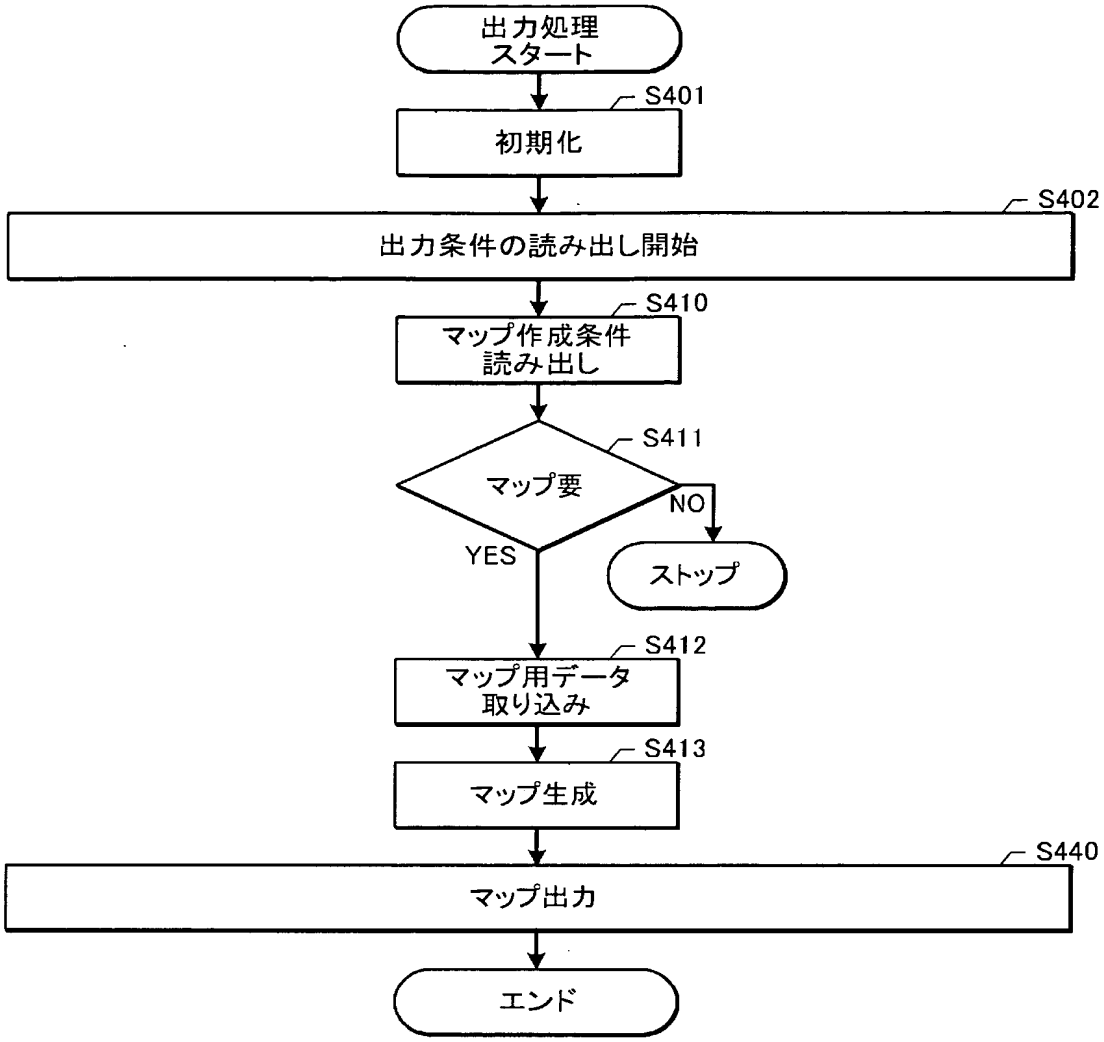
[図33]


2:入力装置
3:記録装置
210
対象企業の文書d条件入力部
230
抽出条件その他入力部
220
比較対象文書群P条件入力部
310
条件記録部
1:処理装置
110
対象企業の文書d読み出し部
120
索引語(d)抽出部
130
比較対象文書群P読み出し部
140
索引語(P)抽出部
142
IDF(P)演算部
320
作業結果格納部
155
同業界企業検索部
160
同類文書群S選出部
170
索引語(S)抽出部
171
IDF(S)演算部
173
中心点算出部
180
文書抽出部
330
文書格納部
4:出力装置
410
マップ作成条件読み出し部
412
マップ用データ取り込み部
440
マップ出力部

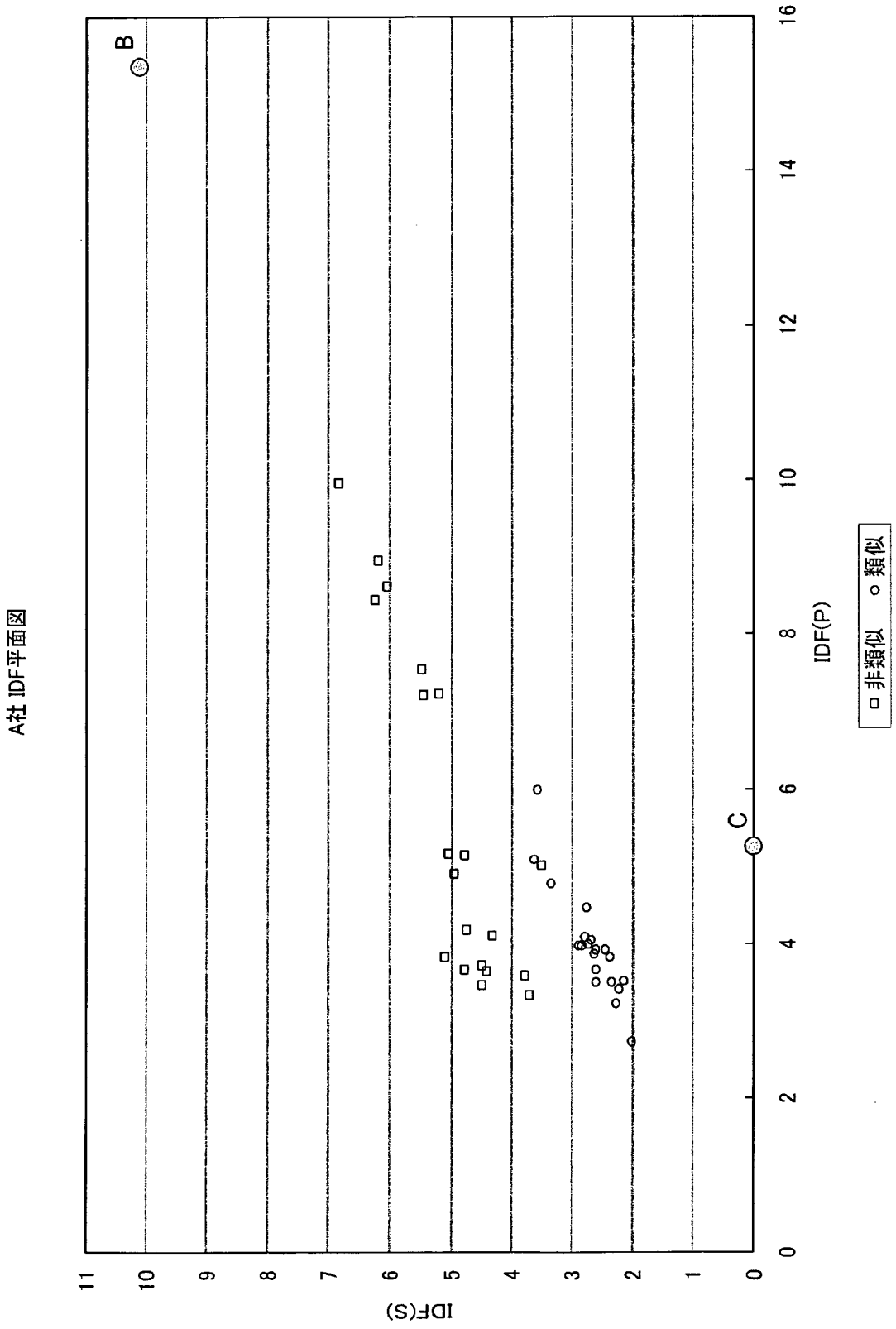
[図34]


中心点算出処理
スタート
S101
初期化
S102
読み出し区別
S110
対象企業の文書d
読み出し
S120
索引語(d)抽出
S130
比較対象文書群P
読み出し
S140
索引語(P)抽出
S142
IDF(P)演算
S155
同業界企業検索
S160
同類文書群S選出
S170
索引語(S)抽出
S171
IDF(S)演算
S173
中心点算出
S180
文書抽出
エンド

[図35]


出力処理
スタート
S401
初期化
S402
出力条件の読み出し開始
S410
マップ作成条件
読み出し
S411
マップ要
NO
YES
ストップ
S412
マップ用データ
取り込み
S413
マップ生成
S440
マップ出力
エンド

[図36]


A社 IDF平面図
IDF(S)
IDF(P)
B
C
□非類似 ○類似
0
1
2
3
4
5
6
7
8
9
10
11
0
2
4
6
8
10
12
14
16

[図37]